IRANY SALGADO MAZZOLA

PROJETO DE DATA WAREHOUSE DIMENSIONAL

FLORIANÓPOLIS - SC

2002

UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Irany Salgado Mazzola

PROJETO DE DATA WAREHOUSE DIMENSIONAL

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação

João Bosco da Mota Alves

Orientador

Florianópolis, julho de 2002.

PROJETO PARA MODELAGEM DE DATA WAREHOSE DIMENSIONAL

Irany Salgado Mazzola

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação na Área de Concentração (Computação Aplicada) e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Fernando Álvaro Ostuni Gauthier, Ph D
Coordenador do Curso de Pós-Graduação

Banca Examinadora

_____

João Bosco da Mota Alves
Orientador
Presidente da Banca

_____

Fernando Álvaro Gauthier
Membro da Banca

_____

Luiz Fernando Jacinto Maia
Membro da Banca

Sumário

Capítulo 7

Capítulo 8

Capítulo 9

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

DBAs – *Data Base Administrators*

DBMS – *Data Base Management System*

DW – *Data Warehouse*

EIS – *Executive Information Systems*

KDD – *Knowledge Discovery in Databases*

MER – *Modelo Entidade-Relacionamento*

MOLAP – *Multidimensional On-Line Analytical Processing*

OLAP – *On-Line Analytical Processing*

OLTP – *On-Line Transaction Processing*

RDBMS – *Relational Data Base Management System*

ROLAP – *Relational On-Line Analytical Processing*

SI – *Sistemas de Informação*

RESUMO

O presente trabalho sugere o projeto de um modelo de Data Warehouse Dimensional a ser utilizado como uma ferramenta tecnológica de análise *on-line* de dados, em processos de tomada de decisão de organizações comerciais.

Para tal estudo foram analisados o projeto e a arquitetura de um DW, o armazenamento de seus dados que pode estar intimamente ligado tanto a bancos de dados dimensionais e relacionais, quanto a data marts, cujos processos se baseiam tradicionalmente em atividades como extração, aquisição, refinamento, armazenamento e acesso a dados.

Os Data Warehouse se utilizam, geralmente, de bancos de dados como depósitos menores. Mas estes, sem um projeto adequado, se mostram mais hábeis em transações *on-line*, com suas aplicações se ajustando com mais eficiência sobre funções rotineiras e operacionais, de caráter periódico curto do que necessariamente para análise histórica de uma determinada situação.

Por esta razão, projetar Data Warehouses têm se apresentado como uma das tarefas mais desafiadoras aos analistas e desenvolvedores de sistemas de informação para empresas. E por mais adequados que possam parecer ao processo decisório, é preciso considerar que sua capacidade de acúmulo histórico de dados de longos períodos (tempo nunca é inferior a três anos), um projeto que obedeça tal fim deve contar com mais consistência e relevância em se tratando de qualidade de dados.

Nos últimos tempos, os especialistas da área de Data Warehouse têm procurado projetá-los sob diversos pontos de vista, procurando estimar sua capacidade nas mais diversas áreas das empresas. Integrar atividades e dados operacionais, administrativos e financeiros é a proposta do que tem se provado como fundamento do Data Warehouse: os data marts.

Portanto, o modelo de DW aqui apresentado vai basear-se no estudo teórico de projetos, esquemas e tipos de modelagens capazes de suportar os processos decisórios com os quais se defrontam as organizações atuais.

ABSTRACT

The present work suggests the project for a model of a Dimensional Data Warehouse, to be used as a technological tool of data analysis, in processes of decision making for commercial organizations.

For such a study, the project and the architecture of a DW were analyzed, as well as the storage of its data wich can be intimately linked to dimensional and/or relational databases both useful to date marts, whose processes are based traditionally on activities as extract, acquisition, cleansing, storage and access to data.

Data Warehouse usually make use of databases as smaller deposits. But these ones, without an appropriate project, are shown more skilled in transactions on-line, with its applications adjusting with more efficiency over routine and operational functions, on short periods than for analysis of long historical periods.

To project a good Data Warehouse model has been one of the most challenging tasks to analysts and developers of information systems. And no matter how appropriate they can seem to the decision making process, it is quite necessary to consider that its requirements of historical accumulation for long periods data (whose time is never under three years), always direct a project to as much consistence and quality of data.

For the past years Data Warehouse specialists have been trying to project models under several point of view, evaluating its ability to deal with information coming up from operational and extenal sources. To integrate all sort of data, executive, operational and financial, has been a complex task and yet, proving data marts one of the most critical basis to Data Warehouse.

Therefore, the model of DW here presented is based on the theoretical study of projects, outlines and types of modellings able to support decision process which are so current for organizations in our days.

INTRODUÇÃO

1.1. Introdução Geral

O ambiente globalizado e altamente competitivo em que estão inseridas as organizações em nossos dias, tem oferecido duas opções de comportamento às mesmas: ou elas sucumbem por não apresentarem condições suficientes de fazerem frente às suas concorrentes; ou criam instrumentos que orientaem suas ações em direção ao sucesso e lhes possibiletem fazer face à acirrada competição de que são prisioneiras.

Orientar as organizações em direção a melhores fatores de vantagens competitivas nos negócios envolve a otimização de recursos, a distribuição de melhores produtos e serviços e a busca constante pelas expectativas do mercado consumidor. Isto inclui, tanto as atividades tradicionais da organização como a concentração sobre as demandas do mercado e nichos de marketing, quanto as novidades trazidas pela engenharia dos negócios, com apelos sobre a aquisição de tecnologia de informação, que juntas podem facilitar, pela competência, o processo de desenvolvimento e integração organizacional.

A combinação estruturada e formalizada de componentes de software e hardware vem sendo usada com o objetivo de adquirir, analisar e relatar, além de fornecer acesso à vasta gama de dados processados pelas organizações. Partindo-se desta definição específica, a Engenharia da Informação tomou a tecnologia do Data Warehouse, como uma das ferramentas mais eficientes de negócios, por permitir a um grande número de usuários finais, gerar consultas dinâmicas a bancos de dados, em processos decisórios dentro das empresas.

Nos últimos anos os conceitos de Engenharia da Informação, Bancos de Dados Relacionais, Data Warehouses, ferramentas OLAP (On-Line Analytical Processing) e OLTP (On-Line Transaction Processing), além dos data marts, surgiram como fórmulas de incalculável valor para solucionar os problemas que envolvem o armazenamento e o acesso às informações dos negócios organizacionais.

1.2. Relevância da Pesquisa

Este trabalho possui como idéia central o modelo de um Data Warehouse, utilizando ferramenta on-line, baseando-se em data marts e investigando a sua eficiência em processos decisórios cujos diagnósticos possam ser vitais para a sobrevivência de uma organização.

Considera-se que os resultados a serem aqui obtidos possam ser úteis tanto à comunidade científica, quanto às empresas que empregam tal tecnologia em seu dia a dia visando, além da própria função de análise, a melhor utilização dos recursos da organização.

A importância do presente trabalho está no fato de que se pretende conciliar os estudos teóricos de pesquisadores como Ralph Kimball e Nigel Pendse, cuja colaboração em termos de critérios visa a representação formal de depósito de dados sob uma série de padrões para processos de suporte à decisões com ferramentas *on-line*.

1.3. Problema a Tratar

O objetivo geral deste trabalho é apresentar um modelo demonstrativo de Data Warehouse de conformidade com regras e padrões que serviram, até agora, como orientação geral para sua própria construção.

Para isso, utilizou-se no estudo, uma versão voltada para a divisão da empresa em *Data Marts*, dos quais as atividades devem corresponder com exatidão às expectativas de seus usuários.

Foram usados também, estudos de diferentes pesquisadores, que embora não sejam idênticos na forma, apresentam a essência que fundamenta ferramentas tecnológicas de tal tipo.

Para alcançar o objetivo geral de formular um modelo de Data Warehouse, faz-se necessária a divisão por etapas de trabalho, destacando-se três objetivos específicos:

1. Divisão da organização em áreas especialmente definidas e atividades que retratem seu histórico de dados.

2. Elaboração de um modelo dimensional de armazenamento de dados para Data Warehouse, com múltiplos recursos, incluindo ferramentas de análise de dados.

3. Apresentação do modelo de Data Warehouse, sobre a teoria da modelagem dimensional, abordagem bottom up e esquema Star.

1.4. Estrutura do Trabalho

O estudo está segmentado em capítulos na seguinte forma:

O primeiro capítulo apresenta uma introdução ao tema do trabalho, procurando mostrar o contexto do armazenamento dos dados dentro de uma organização. Ali, parte-se do princípio que a análise dos dados vem sendo orientada através de ferramentas com tecnologia mais potente tal como a do Data Warehouse, onde é possível combinar e formalizar componentes como os bancos de dados relacionais e/ou dimensionais e as ferramentas de processo de transações e de análise on-line, mais conhecidas como OLTP (On-Line Transaction Processing) e OLAP (On-Line Analytical Processing).

O capítulo dois vai destacar os aspectos mais relevantes da Engenharia da Informação, mostrando seu histórico, conceito, características gerais, papel dos dados e metodologia para sua aquisição, refinamento e acesso.

O capítulo três volta-se para os Sistemas de Informação, de um modo generalizado. Assim é feito quando descreve o conceito de sistema, embora aprofunde-se um pouco mais ao fazer referência ao armazenamento de dados em bancos, até chegar aos Sistemas de Informação Executiva, de onde é introduzida uma visão das informações a partir da perspectiva dos negócios.

No capítulo quatro, são mostrados os processos de transações sobre dados armazenados em sistemas operacionais cuja utilização, apesar de imprescindível para a operacionalização da organização, apresenta características extremamente voláteis quando seu destinados à função de análise.

O capítulo cinco oferece uma visão do Data Warehouse como depósito de dados. Seu histórico em plataformas do tipo mainframe, sua utilização através da Internet e em ambientes Intranet e/ou Cliente/Servidor. Também é apresentada a arquitetura do

mesmo, seus esquemas e abordagens, assim como os modelos de tabelas utilizados para armazenamento de tipos DW e de seus dados.

As aplicações de dados são apresentadas no capítulo seis, onde também são mostrados os modelos mais utilizados de bancos de dados, as estratégias de acionamento dos dados, a partir da utilização da técnica de garimpagem e a utilização da ferramenta OLAP em Data Warehouses.

No sétimo capítulo são tratadas as etapas da modelagem do Data Warehouse, levando-se em consideração a teoria apresentada nos capítulos anteriores e que vem a servir como orientação na proposta do modelo.

No oitavo capítulo, são mostradas as etapas da modelagem do Data Warehouse, considerando algumas áreas da empresa e suas necessidades de informação. Também é apresentada a construção, passo a passo, de tabelas dimensionais, visões e diagramas, como também os resultados, a discussão e a conclusão do capítulo.

No capítulo nove faz-se a conclusão do trabalho como um todo.

E finalmente, no décimo capítulo são relacionadas as referências bibliográficas utilizadas no trabalho.

ENGENHARIA DA INFORMAÇÃO

2.1. A Teoria da Engenharia da Informação

A Engenharia da Informação (EI) orienta a representação das atividades dos sistemas de informação da organização, como forma de prover uma visão geral da situação da mesma. Através dela pode-se criar fundamentos para necessidades de análise da corporação, em termos de sistemas de informação com base em novas tecnologias.

No projeto da Engenharia da Informação são mostrados a missão da organização, seus objetivos gerais e específicos, seus planejamentos estratégico e tático e seus modelos de dados, além dos processos fundamentais às suas operações e bases da tecnologia de computadores que exercem papéis essenciais no que diz respeito ao acesso da informação.

A Engenharia da Informação apresentou, inicialmente, os conceitos de análise de dados e de técnicas de projetos de bancos de dados que poderiam ser usados por administradores de bases de dados (DBAs) e analistas de sistemas, visando o desenvolvimento de projetos e sistemas de dados que fossem baseados na compreensão das necessidades operacionais das organizações dos anos 80.

A partir da década de 80, a EI envolveu-se com uma concepção dirigida para negócios, onde passou-se a considerar o ambiente organizacional sujeito à mudanças vertiginosas. Esta variante do conceito da Engenharia da Informação propiciou às organizações a capacidade de alinhar diretamente seus sistemas de informação com suas estratégias direcionais e objetivos organizacionais, conforme estes foram estabelecidos por seus gerentes e como era de praxe fazê-lo na década de 90. Isto permitiu às corporações construírem sistemas de arquitetura aberta para qualquer que fosse a plataforma de software e hardware, como também para ambientes Host Based ou Cliente/Servidor.

2.2. Histórico da EI

A origem e o conceito de Engenharia da Informação desenvolveram-se a partir de 1976, em Perth, Austrália, pelo pesquisador e também diretor administrativo de informações da Information Engineering Services Pty Ltd., Clive Finkelstein.

Finkelstein, que ficou conhecido como "pai" da Engenharia da Informação, desenvolveu o conceito da mesma em trabalho original na área, onde estabeleceu a conexão entre o planejamento estratégico de negócios de uma empresa e seus sistemas de informação.

Nos anos seguintes à criação da Engenharia da Informação, Finkelstein trabalhou em uma série de seis artigos sobre o tema, que viriam a ser publicados também nos Estados Unidos, em 1981.

Em co-autoria com o pesquisador americano James Martin, Finkelstein publicou, em novembro do mesmo ano, um relatório intitulado *"Engenharia da Informação"*. Estas publicações documentaram a chamada variante orientada para negócios, da engenharia da informação e ambos os autores estenderam paralelamente suas pesquisas nesta área, o lhes tem rendido vários artigos e livros.

Atualmente, as ferramentas da EI se tornaram indispensáveis ao planejamento dos sistemas de informação, a modelagem de dados e a de processos como meio de tradução aos sistemas operantes. E, enquanto crescem a competitividade por maiores fatias de mercado entre as corporações, aumentam as redes organizacionais computadorizadas nas quais as técnicas da Engenharia da Informação são vitais.

2.3. Conceitos da EI

A Engenharia da Informação envolve o conjunto de tarefas e técnicas integradas e orientadas a um plano de negócios, modelagem de dados, processos, projetos e implementação dos sistemas em uma organização. Isto é, de um modo geral, o que vai permitir-lhe maximizar recursos de capital, de pessoal e de informação para alcançar os objetivos corporativos de acordo com a visão dos negócios.

Para Finkelstein (1997), a Engenharia da Informação possui muitos e variados tipos de propósitos, tais como o planejamento organizacional, a re-engenharia do

negócio, o desenvolvimento de aplicações e o planejamento dos sistemas de informação e dos sistemas de re-engenharia, que são resumidos no conceito de EI dado pelo autor de que a Engenharia da Informação é, na verdade, um conjunto integrado e evolutivo de tarefas e de técnicas que direcionam a comunicação na organização, habilitando-a a desenvolver pessoas, procedimentos e sistemas, de modo a alcançar sua visão de negócios.

Martin (1991), definiu a Engenharia da Informação como sendo a aplicação de um conjunto interligado de técnicas formais de planejamento, análise, projeto e construção de sistemas de informação sobre uma organização como um todo ou sobre um de seus principais setores.

O autor acrescenta que, em virtude da complexidade das organizações, as técnicas da Engenharia da Informação não podem ser efetuadas sem o uso de ferramentas automatizadas. Conclui-se, então que as funções de planejamento, análise, projeto e construção, passaram a englobar a Engenharia de Software, mesmo que de uma forma diversa. Martin (1991), também faz referência à EI como um conjunto interligado de técnicas automatizadas no qual são construídos modelos da organização, modelos de dados e modelos de processos em uma abrangente base de conhecimentos, a fim de serem usados para criarem e manterem sistemas de processamento de dados.

Alguns anos depois das primeiras publicações sobre o assunto, Finkelstein (1997), redefiniu o conceito de EI de forma mais abrangente. Ali, seriam utilizados os planos estratégico e tático dos negócios na identificação das informações necessárias aos gerenciadores, visando alcançar e concluir estes planos e seus dados e a partir deles saber de onde a informação é derivada. Pois, para o autor, a EI representava as regras dos negócios, dados e informações em modelos, mas também dos processos de negócios formatados em modelagens específicas. Ainda, estes modelos de dados e processos identificariam claramente os requerimentos de gerenciamento de negócios, por introduzi-los, na prática, sob forma de sistemas em plataformas de hardware ou software usando qualquer linguagem de programação ou ferramenta de desenvolvimento, fossem elas implementadas ou utilizadas como sistemas host-based ou cliente/servidor.

Como dedução simples tem-se que, a Engenharia da Informação utiliza os planejamentos estratégico e tático para identificar os tipos de informações necessárias aos níveis gerenciais da organização e destaca entre os dois tipos de planejamento, a origem dessas informações. Feito isto, as regras do negócio, os dados e as informações contidas nos modelos de dados e os processos de negócios e seus modelos são apresentados, objetivando definir com mais clareza os requisitos da organização em termos de implementação de sistemas, sejam quais forem suas plataformas de hardware e software, tipo de linguagem de programação ou ferramentas de desenvolvimento.

2.4. Características Gerais da EI

De acordo com Martin (1991), embora a Engenharia da Informação não possa ser considerada uma metodologia rígida, mas sim uma classe genérica de metodologias, é preciso que se reconheça suas características peculiares tais como:

- O emprego de técnicas estruturadas em nível de organização.
- O processamento top-down.
- A criação de uma estrutura para desenvolvimento computadorizado da organização.
- A confecção separada dos sistemas da estrutura.
- A participação ativa dos usuários finais.
- A criação de um repositório de conhecimentos sobre a organização envolvendo modelos de dados, de processos e projetos de sistemas de armazenamento, tipo Data Warehouses.
- A facilidade na evolução do sistema, a longo prazo.
- A identificação do pontos que formam e melhoram os objetivos estratégicos da organização.

2.5. Papel dos Dados na EI

Os dados são as figuras centrais da Engenharia da Informação. Eles constituem a unidade básica, a mais elementar de qualquer organização. Através deles, pode-se

compreender o histórico, a estrutura e a perspectiva de cada área da organização. Por esta razão, eles são armazenados e mantidos como recursos para a tomada de decisão.

Cada área da organização possui versões, arquivos e procedimentos particulares com relação a dados. No entanto, pela própria estrutura da organização, essas áreas relacionam-se através de um fluxo complexo de documentos, o que, normalmente, pode levar à incompatibilidade e inflexibilidade nas transações internas da corporação.

Annes (2000), define dados como sendo qualquer elemento identificado em sua forma bruta, que por si só não conduz à compreensão de determinado fato ou situação.

Compreende-se, portanto, que o principal propósito da Engenharia da Informação é coletar, refinar e apresentar os dados de forma trabalhada e utilizável, de modo que os usuários da informação possam compreender o contexto corporativo, através de informações para utilizar com eficiência os recursos de que dispõe a organização.

O conhecimento dos dados trabalhados leva ao conhecimento de fatores organizacionais imprescindíveis a habilitar a empresa a utilizar com eficiência seus recursos disponíveis para alcançar seus objetivos.

Rezende (1997), acrescenta que as técnicas de arrolamento dos dados devem-se basear em quatro fatores, a saber:

- Fatores Estratégicos e Táticos
- Fatores Culturais
- Fatores Econômico e financeiros
- Fatores Operacionais

Finalmente, na Engenharia da Informação é permitido usar técnicas de levantamento de dados semelhantes àquelas usadas, com o mesmo objetivo, na Engenharia de Software. São elas: A observação pessoal, a aplicação de questionário, entrevistas, seminários ou reuniões planejadas, a pesquisa e as técnicas mistas. Esta última tem sido mais empregada, por possibilitar a integração de todas as outras técnicas.

Para Melendez (1990), o mapeamento das necessidades ou levantamento de dados requer que os mesmos sejam vistos a partir das seguintes dimensões ou perspectivas:

- **Estrutura Organizacional** – Dimensão que faz correspondência dos dados com o aspecto humano do sistema geral da organização e que envolve sua estrutura decisória, gerencial e operacional. É onde se divide os dados da organização em áreas e hierarquias que estarão efetivamente envolvidas nos projetos e atividades.

- **Processos** – É a dimensão que representa os dado pelo funcionamento corporativo, através das perspectivas dos procedimentos e das atividades normatizadoras destes. Envolvem as rotinas de trabalho e todas a normas que regulamentam estes processos.

- **Sistemas Aplicativos** – É a dimensão onde os dados e parte dos processos em que se encontram são executados eletronicamente. Esta dimensão está ligada à criação de arquivos de dados e de programas que efetuam todos os procedimentos, tais como cálculos, transações, pedidos de compras, etc. que, em épocas passadas eram feitos manualmente pelo pessoal responsável.

O levantamento dos dados permite que se chegue às conclusões sobre as situações de normalidade e aspectos problemáticos no dia a dia da organização. Através dele, os processos e sistemas individuais de dados podem ser coordenados, permitindo-se que se relacionem de forma adequada.

2.6. Enciclopédia e Dicionário

A enciclopédia é a estrutura básica da Engenharia da Informação. Ela funciona como um repositório ordenado, cumulativo e computadorizado das informações relativas ao planejamento, análise, projeto construção e manutenção de sistemas. Para Martin

(1991), algumas ferramentas informatizadas da EI contém dois tipos de repositórios: O Dicionário – que contém nomes e descrições de tipos de itens de dados, processos, variáveis, etc. e a Enciclopédia – que contém informações sobre o dicionário, além de uma representação completa de planos modelos e projetos, com ferramentas que fazem a verificação cruzada, a análise de correlação e a validação. Em resumo, a enciclopédia armazena o significado em diagramas e garante a consistência desta representação.

A representação gráfica das atividades do sistema de informações pode ser vista sob o molde de uma pirâmide. Deste modo, a EI pode aplicar facilmente as técnicas formais de implementação de sistemas de qualidade em tempo certo.

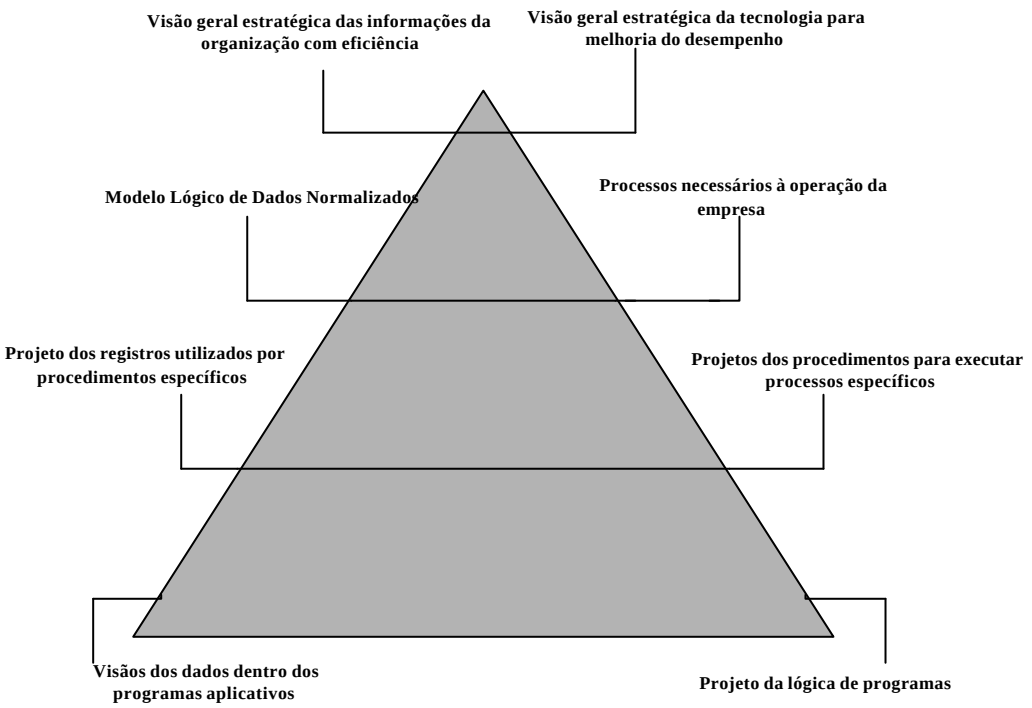

Fig. 2.6. Pirâmide das atividades, informações, modelos e estratégias organizacionais

A pirâmide, cuja estrutura das atividades e informações concernentes à organização deve ser orientada e dividida em camadas, conforme mostrada acima.

No topo, ou primeira camada, estão concentradas as funções de planejamento, referentes ao planejamento estratégico de informações necessárias na condução do

negócio, com o máximo de eficiência. Nesta camada estão contidas as informações tais como a missão, objetivos, fatores de sucesso e fatores críticos, assim como a tecnologia a ser usada para melhorar a performance da organização. Ou seja, aqui, cria-se uma visão geral e de alto nível da empresa, onde figuram suas funções, dados, processos e necessidades de informação.

Na camada seguinte, também chamada de camada de análise, estão contidos os modelos lógicos dos dados e processos fundamentais necessários ao funcionamento da organização. Aqui, feito o levantamento dos dados, seus processos e relacionamentos, vai-se projetar o modo como serão integrados à operacionalização das áreas do negócio.

A terceira camada representa o projeto dos registros utilizados, em procedimentos específicos e a forma como serão implementados. Na construção desta camada, sugere-se a participação direta do usuário final, como fim de melhorar a qualidade de apresentação do sistema.

Finalmente, na base da pirâmide, ou camada de construção, estão a visão dos dados dentro do programa aplicativo, assim como o detalhamento ou a entrada para um gerador de programas e implementação dos procedimentos propostos. Aqui utiliza-se linguagens de alto nível e ferramentas avançadas e de um modo geral, quase sempre procura-se vincular o projeto a um protótipo.

2.7. Diagramas

Diagramas são técnicas gráficas muito utilizadas na Engenharia da Informação. Isto se deve ao fato de que um dos trabalhos mais difíceis é a visualização dos mecanismos que são projetados em programas e sistemas.

Nos últimos tempos tem havido várias tentativas de se projetar programas diagramaticamente e usar editores gráficos em programação. Muito embora os diagramas venham se aperfeiçoado à medida que diferentes técnicas estruturadas são desenvolvidas ou atualizadas, nenhuma técnica gráfica tem se mostrado satisfatória. Aparentemente, isto se deve ao fato de que os projetos de programas mostram-se tão complexos que quase sempre chegam ao ponto de requererem mais do que um só tipo de

diagrama, de modo que o projetista é incapaz de visualizar diferentes aspectos do projeto como um todo.

A construção de um diagrama visa oferecer uma visão geral dos dados da organização ou de parte dela, selecionada para estudo e onde irão figurar os objetos e seus relacionamentos.

Os diagramas podem compor um projeto na Engenharia da Informação. A maioria deles refere-se aos dados contidos no modelo de dados, cujas informações devem ser interligadas. Alguns tipos são relacionados abaixo:

- Diagrama de Ação
- Diagrama de Fluxo de Dados ou Diagrama de Dependência
- Diagrama de Árvore e Tabelas de Decisão
- Diagrama de Estado Limitado

Diagramas são, quase sempre, gerados a partir de uma enciclopédia que contém as informações globais sobre o projeto. São unidos por ferramentas computadorizadas de diagramação capazes de representá-los como se fossem uma faceta do projeto, podendo ser exibidos nas várias janelas nas tela de uma estação de trabalho.

Em condições normais, os vários diagramas de um projeto são ligados logicamente em um Hiperdiagrama. A atualização de do projeto pode se dar através da modificação de um diagrama, seguindo-se da alteração automática de todos os outros.

2.8. Metodologia da EI

A Engenharia da Informação envolve dois fatores básicos em sua metodologia: uma documentação textual completa e uma descrição dos conjuntos de procedimentos, que podem ser representados gráfica e detalhadamente, das tarefas a serem aplicadas a esta metodologia. Tudo isso objetiva servir de apoio aos procedimentos documentais e diagramas técnicos, assim como para os exemplos práticos de aplicações específicas da metodologia.

De acordo com Finkelstein(2000), a metodologia da EI que pode ser aplicada na organização apoia-se sobre os seguintes fatores:

- **Planejamento do Negócio** – O objetivo é aplicar o *estado da arte* aos negócios, visando a análise lógica de seu modelo enquanto se emprega técnicas para planos mais produtivos, consistentes e coerentes na empresa. No modelo do negócio estão representados o plano da empresa, sua estrutura organizacional, regras básicas de negócios, inovações nas atividades que podem incrementar a produtividade, a qualidade e a eficiência, análise do ambiente externo, fatores críticos de sucesso e fracasso, o estabelecimento de objetivos e de alternativas que vão determinar novas direções e oportunidades para a empresa.

- **Modelagem de Dados** – É a representação abstrata dos dados, informações e conhecimento que a organização detém. O modelo de dados é desenvolvido baseado no plano (seja como um todo, seja parcialmente) dos negócios da empresa, desenvolvendo-se um esquema de banco de dados. Cabe à organização desenvolver um modelo integrado dos dados históricos, com diferentes aplicações e expandi-lo para a organização inteira. A modelagem de dados é usada para identificar, exibir e controlar os subconjuntos de modelos de dados funcionais da organização e suas perspectivas de utilização.

- **Modelagem de Processos** – É a representação abstrata da atividade independente da tecnologia. Utilizando-se uma abordagem retrospectiva, um modelo de processo é desenvolvido e rigorosamente baseado no modelo de dados, afim de demonstrar a aquisição e a disseminação dos dados, das informações e do conhecimento. Os procedimentos manuais e automatizados, de um ou mais sistemas, devem ser mostrados, assim como os projetos de suas aplicações, que proverão a base necessária à modelagem de processos. A Engenharia da Informação também reutiliza dados e processos lógicos,

capacitando a modelagem de processos a facilitar a re-engenharia dos negócios.

- **Projetos de Sistemas** – O objetivo principal dos projetos de sistemas é criar ou documentar detalhadamente as especificações de um sistema. Utilizando-se a abordagem retroativa de engenharia, o projeto do sistema é rigidamente baseado nos sub-conjuntos de modelos de dado e de processos, que sejam mais relevantes para o sistema. Os requerimentos de tecnologia, configuração e capacidade de tarefas projetadas variam em função do tipo e da situação específica da organização. Os projetos de sistemas têm por objetivo otimizar o modelo de negócio em sua implementação, dadas a sua capacidade e seleção de tecnologia. Projetos de sistemas que usam abordagem retroativa de engenharia podem ser desenvolvidos para documentar *legacy systems* e incorporá-los na arquitetura da Engenharia da Informação, ligando-os, apropriadamente aos componentes de modelos de negócios. Os projetos de sistemas incluem distribuição, mecanismos de acesso, interfaces de usuários e sistemas e configuração detalhada da tecnologia.

- **Implementação de Sistemas** – A implementação dos sistemas visa transforma o projeto todo em um sistema concreto de informações. Todos os aspectos de instalação, construção, conversão de dados, testes, documentação, treinamento e transição devem ser coordenados com a evolução da infra-estrutura necessária para alcançar o sucesso da implementação. Com freqüência, a aquisição e a instalação de pacotes de hardware e software precedem a compleição do projeto de sistemas, para que as novas tecnologias possam ser implantadas sem causar atrasos no desenvolvimento.

Melendez (1990), propõe uma metodologia apoiada sob oito premissas básicas e integradas, que podem ajudar a definir e processar os dados na EI. São elas:

- **Planejamento Estratégico de Sistemas de informação** – Premissa voltada para a formalização dos objetivos e fatores de sucesso, ambos críticos, apontados pela alta administração. Inclui a modelagem da organização em si e o planejamento estratégico das informações e suas aplicações. Deve identificar formas de melhorar a estrutura organizacional e o uso otimizado da tecnologia na organização

- **Projeto Centrado em Dados** – Diz respeito ao conjunto de técnicas voltado para a administração e a modelagem de dados formais. Esta última, buscando simplificar e reduzir custos na construção e manutenção dos sistemas.

- **Métodos Novos e Técnicas** – Através da substituição de técnicas estruturadas convencionais e de metodologias manuais por aquelas computadorizadas, mais velozes e capazes de, por exemplo, efetuar verificações cruzadas abrangentes através de um sistema complexo.

- **Computação ao Usuário Final** – O desenvolvimento de sessões de projeto aplicativo aos usuários, orientado pelos analistas, objetivando especificar com maior precisão os sistemas a serem projetados. Na prática da criação de protótipos a utilização das críticas dos usuários no desenvolvimento dos sistemas tem sido bastante utilizada.

- **Projeto Automatizado** – Construção de diagramas contendo todos os aspectos da engenharia de sistemas e mostrados em estações de trabalho, que disponibilizarão o projeto como um todo. As ferramentas CASE são a base dos geradores de programas destinadas a esta tarefa.

- **Produtividade no Processamento dos Dados** – Estabelecimento de uma associação entre as ferramentas de automação de projeto e os geradores de programas pode incrementar a velocidade e a qualidade da construção de sistemas.

- **Reutilização de Projetos e Programas** – A utilização da abordagem *top-down* da engenharia da informação, pode ajudar a identificar aqueles processos utilizados muitas vezes pela organização. A reutilização do todo ou de partes de projetos e programas pode reduzir custos e tempo.

- **Sistemas Especialistas** – A EI usa sistemas especialistas para ajudar os planejadores, analistas e projetistas a criarem melhores sistemas, visto que os mesmos aplicam o processo de inferência a uma base de conhecimentos contendo dados e regras, a fim de fazer com que o computador simule o raciocínio humano, alcançando, algumas vezes, maior precisão do que este último.

Como mencionado anteriormente, em uma organização, os modelos de dados são estruturados em torno das áreas específicas, às quais estarão associados arquivos e procedimentos próprios e modelos de processos característicos e reflexivos das áreas para as quais foram desenvolvidos. Mesmo assim, todo os conteúdo destes fluxos complexos de documentos devem possuir alguma forma de integração, visto que existe grande possibilidade de diferentes áreas possuírem os mesmos dados, levando o sistema de informações a registrar casos de redundância e/ou incompatibilidade entre os mesmos, causados por diferentes versões. Por esta razão, indiferente à metodologia a ser utilizada, a EI procura criar planos e modelos de alto nível, onde os sistemas a serem construídos são vinculados, para evitar que se tornem incompatíveis.

2.9. Utilização de Projetos

Projetos da Engenharia da Informação são abrangentes. Eles envolvem diversas e diferentes áreas de pesquisa que devem integrar-se, principalmente, através de ferramentas computadorizadas.

A experiência mostra que os projetos automatizados da EI resultam em produtividade mais elevada, pois provêem a empresa com ferramentas que coordenam seus planos estratégicos, operacionais e organizacionais de demanda.

Seguindo os objetivos de implantar a Engenharia da Informação nas empresas, houve uma mudança paralela quanto à cultura do tipo de sistemas de informação que a suportariam. Sob este aspecto, as inovações construtivas das ferramentas baseadas na informática, tais como o desenvolvimento de sistemas automatizados de informações, uso intensivo da Internet e Intranets, dos Bancos de Dados Relacionais baseados em processos de transações on-line e, principalmente, do Data Warehouses baseados em ferramentas de análise on-line, ajudaram a alavancar vantagens competitivas para organizações. Compreendeu-se, em determinado momento, que os processos produtivos de bens e serviços caminhariam por si próprios se a empresa estivesse calcada sobre sólidas bases de informações, que lhes serviriam também de mapas para melhores oportunidades no mercado. A Engenharia da Informação cresceu em importância à medida que a tecnologia se tornou o principal pilar de sustentação de processos produtivos e administrativos. Além disso, tendo toda a sua orientação voltada para os negócios, foi possível direcioná-la de encontro a fatos que modificaram a economia mundial, tal como a globalização. No próximo capítulo, será visto como teorias sobre tecnologia e sistemas de informações têm sido misturados aos processos econômicos de todas as nações do planeta.

SISTEMAS DE INFORMAÇÃO

Tecnologia, economia e sistemas são vocábulos comumente associados em nossos dias para explicar uma série de fenômenos que vêm modificado o panorama mundial. A conexão entre estas diferentes áreas só poderá ser compreendida se for percebida como sendo o conjunto que proporcionou equilíbrio entre fatores sociais, econômicos, políticos e financeiros.

O presente capítulo objetiva mostrar como o mix de elementos com objetivos diferenciados foi propulsor de mudanças radicais na rotina de organizações em diferentes nações, tornando-as idênticas em quase todos os seus procedimentos.

3.1. Tecnologia, Economia Global e Negócios

Três fatores foram fundamentais para a aplicação de tecnologia de informação a negócios:

- A globalização da economia
- A evolução tecnológica
- A re-engenharia dos negócios da empresa

O primeiro fator, ou a chamada emergente economia globalizada, provocou a expansão da mercantilização de produtos, a custos acessíveis para países com legislação menos restritiva. A queda do comunismo e a liberalização das economias asiática e sul-americana apresentaram-se como desafiadores segmentos de mercado, embora suas políticas econômicas instáveis não incentivem o estabelecimento do setor da manufaturaria pelos países desenvolvidos. Por outro lado, a oferta de novas oportunidades nos negócios sinalizou para uma postura de entrada virtual e de valorização da análise da informação no mercado internacional, que é o que a tecnologia de Data Warehouse vem sustentando de maneira eficiente.

O segundo fator a ser considerado é a própria evolução da tecnologia que foi alcançada por alguns países, transformando-os em potências econômicas de informação e conhecimento.

Esta evolução, que começou na virada do século XX, foi impulsionada pelas invenção de máquinas com capacidade para realizar o trabalho humano, reduzindo-o em tempo e custo. Ela acabou também, por redirecionar a força de trabalho do homem do campo para a cidade, transformando trabalhadores do setor primário em empregados que exercem atividades relacionadas à educação, saúde, comércio e investimentos financeiros. Estes novos empregos passaram gradualmente a envolver a criação, gerência e distribuição de conhecimento, que foi o que, de fato, impulsionou ainda mais o desenvolvimento da tecnologia aplicada a negócios.

Por último, a transformação do ambiente de trabalho, antes organizado de modo centralizado e hierárquico e rigidamente baseado nos procedimentos operacionais das produções em larga escala de bens e serviços em uma estrutura mais flexível e com nova visão e métodos de trabalho. Neste aspecto, o uso da tecnologia teve papel fundamental. Primeiro, porque ficou comprovado que a tecnologia aplicada à produção em massa podia se auto gerenciar, dispensando a interferência humana. E segundo, porque as os novos estilos de organizações que vieram a ser aplicados às empresas passaram a depender cada vez mais do conhecimento, da capacidade de aprendizagem e do poder de decisão de seu pessoal.

3.2. Sistemas em Conceitos

De acordo com Bio (1985), a partir da metade do século passado o desenvolvimento do conhecimento humano foi acelerado e passou a exigir especialização, provocando também uma necessidade crescente de profissionais capazes de sintetizar as complexidades dos sistemas, relacionando suas partes com o todo.

Bio (1996), conceitua sistema como sendo um conjunto de elementos interdependentes, ou um todo organizado, ou partes que interagem formando um todo unitário e complexo.

Com o passar do tempo, surgiram inúmeros tipos de sistemas, conceitos e aplicações que ao evoluírem, chegaram ao que se conhece hoje em dia. A maior parte destes se desenvolveu após a Segunda Guerra Mundial, obrigando o refinamento de técnicas e instrumentos capazes de explicá-los e representá-los. Atualmente, é comum

ouvir falar de sistemas de defesa, sistemas sociais, sistemas econômicos e sistemas políticos, entre muitos outros.

Por definição, os sistemas comportam-se de modo que suas várias faces precisam ser relacionadas entre si para oferecerem um retrato fiel de suas características, interdependência e funcionamento.

3.2.1. Sistemas Abertos e Sistemas Fechados

Os sistemas, no que concerne seus tipos podem ser abertos ou fechados. Bio (1996), faz a distinção entre os dois, destacando que ser preciso distinguir sistemas fechados, como as máquinas, o relatório, etc., dos sistemas abertos, como sistemas biológicos e sociais: o homem, a organização, a sociedade.

Os sistemas abertos são aqueles que permitem trocas com o ambiente externo, contrariamente aos sistemas fechados do qual nenhuma matéria pode entrar ou sair Bertalanffy (1976).

De um modo geral, são os tipos abertos que reproduzem melhor a idéia sobre o conceito de sistema, pois mostram ao mesmo tempo, o dinamismo e a interdependência das partes com o inteiro, que é orientado para um fim determinado e deve se mostrar em constante interação com o ambiente externo.

O conceito de sistema na Administração retrata as empresas como sistemas abertos. As produções de bens e serviços executados por diversas áreas, pressupõem, ao mesmo tempo, que estas áreas interligadas e interdependentes respondem como um todo às pressões do ambiente externo. Para coordenar de maneia eficaz os esforços da organização a fim de sobreviver, os sistemas de informação têm sido usados há muito tempo como ferramenta de colaboração e de reprodução de relatórios no trabalho.

A seguir, são mostrados conceitos de sistemas de informação, seus componentes, atividades e funcionamento.

3.2.2. Sistemas Informais e Sistemas Formais

Os sistemas de informação podem ser informais ou formais. Os sistemas informais não se apoiam em regras pré-estabelecidas, compromissos ou acordos. Um exemplo destes sistemas são as rede de comunicação entre funcionários. Nelas as

pessoas contam fatos tipo “fofoca de escritório” e outros fatos menos importantes, dos quais não há necessidade de armazenamento mesmo que sejam essenciais para a vida da organização.

Os sistemas formais, por outro lado, são aqueles com definições, dados, regras e procedimentos fixos e praticamente imutáveis.

Os sistemas de informação formais podem, ainda, ser manuais ou computadorizados. Os sistemas manuais usam enormes quantidades de papéis, fichários, pastas e canetas. Seu armazenamento ocupa grandes espaços físicos e sua atualização é lenta.

Já os sistemas de informação formais e baseados no uso de computadores contam com a tecnologia de software e hardware para auxiliá-los nas tarefas de armazenar e disseminar informações. O espaço físico para guardar dados reduz-se ao disco rígido dos computadores. A questão central neste caso é a adequação de programas e equipamentos, além, do treinamento a ser ministrado aos usuários finais.

3.3. Sistemas de Informações

O gerenciamento de informações através de sistemas tem gerado inúmeras definições por parte dos profissionais desta área em permanente desenvolvimento.

Pela definição de Laudon (1998), um sistema de informações pode ser definido tecnicamente como um conjunto de componentes interrelacionados que coleta (ou recupera), processa, armazena e distribui informação para apoiar a tomada de decisão e controlar a organização.

Bio (1985), apresenta outro conceito para o mesmo tema, no qual trata sistema de informações como um subsistema do “sistema empresa”.

Os sistemas de informação podem então, serem entendidos com termos como procedimentos, normas, relatórios, políticas, métodos e processamentos, embora o ponto em comum em qualquer que seja sua definição corresponda à existência de partes distintas e funcionais, relacionadas e interdependentes que vão representá-lo como um todo, fazendo-o funcionar eficaz e eficientemente através da troca de informações.

Sistemas contém subsistemas, que são menores e mais detalhados. Bio (1985), afirma que o chamado sistema empresa pode conter vários tipos de subsistemas tais como o de orçamento, custos, contabilidade, vendas, produção, materiais, marketing, etc.

Os subsistemas de um sistemas não obedecem uma classificação rígida. Mas em geral, eles estão ligados à divisão por áreas de uma organização. A partir da perspectiva gerencial da mesma, eles poderão dar apoio às operações do dia a dia ou a seus processos decisórios.

Dentro dos subsistemas podem existir outros subsistemas menores. Por exemplo, o subsistema de orçamento pode vir a conter um subsistema chamado Fluxo de Caixa, onde vão estar os dados sobre o volume de dinheiro a ser movimentado diariamente pelo caixa da empresa, através de operações de pagamento de dívidas e recebimento de valores correspondentes à venda de mercadorias ou prestação de serviços. Este subsistema registra apenas as operações relacionadas com as atividades acima mencionadas e as atualiza todos os dias.

As informações resultantes da coleta de dados mencionado no exemplo do fluxo de caixa constituem parte de um subsistema.

3.4. Dados e Informações em SI

Antes que um sistema de informações possa corresponder à vasta gama de necessidades de informações no processo de tomada de decisão, o trabalho administrativo deve ser organizado e todos os dados resultantes de ações repetitivas na organização devem estar interligados através de procedimentos.

Os procedimentos definem a ação requerida, quem e quando executa a executa e, feito o registro de seus ciclos completos, eles também fornecem uma espécie de processamento periódico dos dados que mais tarde vão constituir as informações que são os componentes fundamentais de um sistema. Aqui, a informação é entendida como sendo o conjunto de dados coletados sobre eventos significativos para a organização que deverão ser analisados e formatados, usando-se técnicas científicas próprias.

Para virem a ser usados pelo pessoal da organização, os dados, que são os correspondentes primários dos fatos, só poderão se tornar informações quando forem devidamente organizados e tabulados. O processamento necessário para a sua conversão consiste em uma série de operações ou atividades que começam em sua captura, passam pela sua manutenção permanente e culminam com a sua saída.

Estas atividades de manipulação dos dados são, segundo Laudon (1999):

- **Input** – Atividade de captura ou coleta de dados primários dentro ou fora do ambiente da organização.
- **Processamento** – Atividade de conversão dos dados primários coletados através do input.
- **Output** – Atividade de transferência dos dados processados, agora chamados de informação, às pessoas ou atividades onde serão usados.
- **Feedback** – Atividade de avaliação que retorna as informações apresentadas pelo output para colaboradores da corporação, visando corrigir possíveis falhas no input dos dados.

Para exemplo prático, observe-se o rotineiro caso do recebimento, pelo depósito de uma empresa, de matéria-prima com nota fiscal. Apesar das ações de descarregamento e estoque do material se passarem no plano físico, o registro de sua nota fiscal contendo tipo, quantidade e preço de sua aquisição, constitui o dado a ser transformado em informação para alimentar o sistema em um plano mais elevado. Esta informação, que talvez não seja muito relevante do ponto de vista gerencial, pode se tornar fundamental quando analisada em conjunto com outras informações do mesmo tipo, no processo decisório relativo à melhor estratégia para aumentar a rotatividade de materiais usados pela empresa e diminuir custos com estoque.

Portanto, embora os registros dos dados correspondentes à ações operacionais repetitivas não tenham papel como informação significativa para processos de tomada de decisão, é sempre necessário que recebam tratamento adequado para virem a ser, posteriormente, úteis ao sistema de informações.

3.5. Bancos de Dados Relacionais

A abordagem estruturada dos dados dos sistemas de informação tem sido feita, principalmente, sob forma de bancos de dados instituídos sobre coleções de centenas ou milhares de dados armazenados com determinado fim.

Embora a utilização de banco de dados possa tornar mais complexa a estrutura do sistema de informações, ela também é capaz de simplificar o fluxo das mesmas, por diminuir o volume de papéis trocados entre áreas da organização.

Nos informação atuais, a maioria dos bancos de dados utilizada é do tipo Sistema Relacional. Estes, embora possuam uma perspectiva histórica relativamente nova, estabeleceram-se como o principal modelo de dados para aplicativos de processamento de dados comerciais, sendo também aplicados a projetos assistidos por computadores e a outros ambientes.

O modelo relacional de banco de dados faz parte do modelo lógico baseado em registros. Este modelo é assim chamado por ter seu banco de dados estruturado em registros de formato fixo de diversos tipos. Os registros, por sua vez, definem um número fixo de campo, ou atributo, cujo tamanho, também fixo, vem a simplificar o nível de implementação física do banco de dados.

Um banco de dado relacional tira seu nome a partir da existência de uma correspondência íntima entre o conceito de tabela (na qual ele se baseia) e o conceito matemático de relação. As coleções de tabelas que estruturam o banco de dados relacional, contém dados e relacionamento entre dados que são designados por colunas que possuem, cada uma, um nome único. As linhas dessas tabelas, por outro lado, representam um relacionamento entre um conjunto de valores.

As principais características dos bancos de dados relacionais são:

- Compartilhamento de informações comuns
- Efetividade na manutenção e atualização dos dados
- Estruturação efetiva de dados armazenados
- Indexação em qualquer campo
- Relacionamento de dados de bases diferentes através de campo idêntico

- Eliminação de redundância de dados
- Consultas on-line
- Implementação de novos campos ou aplicações
- Adaptabilidade à mudanças de hardware.

Existe sempre a preocupação que a estrutura de um banco de dados não se torne muito complexa ao ponto de apresentar dificuldades de visualização. Por esta razão, os projetistas de bancos de dados criam as visões quando trabalham na tela.

Para Martin (1991), uma visão de um banco de dados é uma representação dos dados percebida por uma pessoa ou por um programa.

As visões são subconjuntos da estrutura global do banco de dados. Elas são parte de uma representação maior que é a enciclopédia central e podem ter nomes que serão mencionados no índice da enciclopédia do sistema de informações, para que possam ser rapidamente recuperadas

3.6. Sistemas de Informação Executiva

A partir da perspectiva dos negócios, um sistema de informações apresenta-se como solução sob dois aspectos: o organizacional e o de gerenciamento Laudon (1998). Estes aspectos enfatizam o conhecimento da organização como um todo para estabelecer, no futuro, a natureza das informações que vão compor seus sistemas de informação.

A chave para o conhecimento da empresa na composição de seus sistemas de informação está em sua estrutura, procedimentos operacionais, política, cultura e pessoal.

Segundo Kelly (1999), os sistemas de informação podem ser tradicionais ou executivos (Executive Information Systems), que diferem dos primeiros pelos seguintes aspectos:

- São especificamente projetados para suprir as necessidade de informações do executivos

- Podem acessar dado sobre assuntos e problemas específicos, bem como agregá-los em relatórios
- Provêem extensas ferramentas de análise on-line, incluindo um direcionamento para análise, relatórios das exceções e capacidade de drill-down
- Acessar uma vasta gama de dados internos e externos
- São particularmente fáceis de usar (dispositivos típicos como mouse e touchscreen)
- Podem ser usados pelos executivos, sem auxílio especializado
- Apresentam informações na forma gráfica

Sistemas de Informação Executiva são ferramentas que oferecem acesso direto e on-line às informações relevantes em formato navegável. Como relevantes, entende-se que as informações sejam acuradas, acionáveis e temporais e que repousem sobre os aspectos da organização de interesse dos executivos e gerentes. A navegabilidade do formato significa que o sistema é especificamente projetado para ser usado por indivíduos com limites em termos de tempo, habilidade e experiência na utilização de computadores.

Os Sistemas de Informação Executiva devem proporcionar aos "tomadores" de decisão facilidade na identificação de estratégias a serem adotadas baseando-as na utilização de dados além, é claro, da possibilidade de exploração das informações para achar fatores problemáticos determinando prováveis soluções.

3.7. Condução dos Sistemas Executivos

Como teoria os sistemas de informações preencheram uma lacuna importante na literatura técnica para a condução executiva e gerencial. A partir deles foram integrados os diversos segmentos da organização que com seus dados diferenciados e fragmentados, não achavam o compasso bem marcado para suas atividades.

Por outro lado, estes processos sem a devida automatização não conseguiriam integrar-se formalmente. Foi precisamente aí que a tecnologia interferiu com a

introdução de sistemas on-line, que vieram a facilitar a coordenação e a apreciação das informações disponíveis. No próximo capítulo serão descritos os processos on-line e os serviços que prestam tanto nas transações operacionais quanto na análise de negócios pelos usuários das empresas.

PROCESSOS ON-LINE

4.1. Tendências da Tecnologia

Nos anos 70, a indústria tecnológica voltava sua atenção para o hardware e seus custos. A década seguinte trouxe preocupações dirigidas aos softwares, tanto como elemento promissor da tecnologia da informação quanto como fonte de vantajosas aplicações. Finalmente, no fim do século passado as corporações partiram para o reconhecimento e a exploração maciços e o gerenciamento de dados para incrementar o atendimento ao cliente, a cooperação com os fornecedores e alavancar vantagens em relação aos competidores.

De acordo com Boar (1998), as organizações atuais estão ocupadas em desenvolver estratégias que as mantenham competitivas. O desenvolvimento de estratégias, para o autor significa construir, compor e manter vantagens sobre os concorrente. Por isso, o número de aplicações voltadas para negócios cresceu de modo quase que incontrolável, inserindo, neste contexto, o Data Warehouse como tecnologia das mais competitivas. Esta última, além de ir ao encontro da necessidade fundamental de competição entre organizações, também pode ser vista como um modo superior de pensamento estratégico na dimensão do tempo. Quer dizer, a tecnologia do Data Warehouse pode ajudar a organização a tirar mais vantagem da base de conhecimento criada por ela própria.

4.2. Tipos de Aplicações para Dados

Existem dois tipos básicos de aplicações da tecnologia da informação para negócios: O primeiro funciona sobre as atividades operacionais rotineiras, de caráter diário, semanal, mensal, trimestral, etc. da empresa. São tipicamente armazenadas, recuperadas e atualizadas por sistemas do tipo *on-line*. Seus dados são incessantes e em caso de interrupção em seu funcionamento, a organização, literalmente, pára de operar. Os sistemas operacionais se enquadram neste modelo e são baseados em ferramentas do tipo OLTP (On-Line Transaction Processing).

O segundo tipo de aplicação é passado sobre a análise dos negócios da organização e gira em torno da interpretação baseada em informações e emissão de relatórios para decidir sobre as ações a serem tomadas no futuro. Estas aplicações são suportadas por dados históricos e sua interrupção não compromete, de imediato, as operações da organização, embora obviamente, acarrete perdas em termos de competitividade. O Data Warehouse contém, em sua maior parte, o segundo tipo de aplicações e sua base se apoia em ferramentas OLAP (On-Line Analytical Processing).

4.3. Sistemas Operacionais

Sistemas operacionais são aqueles que, como o próprio nome sugere, ajudam a organização a operar no seu dia a dia. São a coluna dorsal da empresa, pois os dados contidos neles declaram a maior parte das atividades rotineiras de diversas áreas.

Através dos anos, os sistemas operacionais expandiram-se de tal modo que tiveram que ser redesenhados para se tornarem integrados às funções executadas diariamente nas empresas. Como resultado, acabaram por tornar o funcionamento das mesmas inteiramente dependente de seus dados.

Por trabalharem com dados dinâmicos e com alto nível de detalhamento, os sistemas operacionais tendem a possuir centenas de megabytes, ou até mesmo, de terabytes, em tamanho. Como conseqüência, a consistência, a recuperação dos dados, assim como a performance (que maximiza a transação e minimiza os conflitos de concorrência) são fatores críticos. Por isso, requerimentos com relação ao desempenho e à confiabilidade tornam inadequado o seu uso como suporte à decisões.

Devido à sua importância para a corporação eles sempre são a primeira parte do sistema organizacional a ser ordenada e computadorizada e, via de regra, utilizam os sistemas do tipo OLTP (On-Line Transaction Processing) para efetuar suas operações.

Alguns exemplos de sistemas operacionais mantidos por OLTP são sistemas de reservas de passagens, aplicações contábeis e pedidos de mercadorias.

Os tipos de dados que compõem os sistemas operacionais baseados em OLTP são sempre estruturados e repetitivos e constituem transações curtas, isoladas e atômicas,

que requerem grande especificação e constante atualização. Possuem esquemas de normalização e promovem o volume na taxa de envio e o limite de redundância de dados

Por apresentarem tais qualidades, os dados operacionais se mostram incapazes de funcionar como repositório de fatos e dados históricos para análise de negócios.

As características básicas dos dados dos sistemas OLTP são:

- Atualização constante através de transações on-line
- Consistência não-histórica, ou seja com data não superior a três ou seis meses
- Otimização em função do processo transacional
- Normalização através de bancos de dados relacionais, objetivando facilitar sua atualização, manutenção e integridade

As distinções peculiares a estes dados dificultam respostas rápidas à consultas para fins de análise, ao mesmo tempo que tornam impossíveis as recuperações dos mesmos em tempo mínimo. Para efeito de análise os dados são inconsistentes, mutáveis e com grande quantidade de entradas duplicadas tornando irrealizável a acumulação histórica de dados, porque, basicamente, um sistema OLTP oferece apenas enormes volumes de dados não tratados ou dificilmente inteligíveis.

4.3. On-Line Transaction Processing

Os sistemas operacionais registram os processos essenciais da empresa através de transações on-line. Compreende-se então, que as tarefas que movem as engrenagens da organização, tais como pedidos de mercadoria, fluxo de caixa, operações financeiras e registros de clientes, entre outras coisas, são consideradas transações do tipo OLTP (On-Line Transaction Processing).

Um sistema OLTP é capaz de processar, por dia, milhares de transações com pequenas porções de dados. Seus usuários lidam com um registro por vez e executam as mesmas tarefas inúmeras vezes. Portanto, a maior parte dos relatório gerados em sistemas OLTP utiliza o Modelo Entidade-Relacionamento (MER), que divide os dados

em várias entidades distintas e os transforma em tabelas. O resultado é apresentado sob forma de listagens de tabelas com um diagrama extremamente complexo, que visam supervisionar as atividades essenciais da empresa, mas tornam-se inaceitáveis como instrumento de análise.

Entretanto, como para este tipo de sistema o desempenho e a confiabilidade são fatores fundamentais, não é permitido qualquer atividade como as de pesquisa de dados que, obrigatoriamente, lhe acarretam lentidão.

Um dos exemplos práticos e mais populares de aplicação de OLTP é o sistema de reserva de passagens para companhias aéreas. Neste caso, é permitido aos agentes de viagens efetuarem reserva de lugar, completar a transação de compra da passagem e registrar a operação completa em um banco de dados operacional on-line. Entretanto, até o instante da partida, a contagem dos passageiros permanece instável, o que leva à conclusão que este tipo de informação não é útil na tomada de decisões, embora seja crítico para a função operacional a que se destina.

As aplicações dos sistemas operacionais suportadas por OLTP possuem os seguintes atributos:

- Possuem a responsabilidade pelo registro do trabalho “pesado” executado, através das transações nos sistemas
- Provêem serviços durante 24 horas por dia, em 7 dias da semana, com adequado gerenciamento dos períodos de interrupção
- Priorizam integridade e disponibilidade do banco de dados, que em caso de falha, deve ser recuperado em um período mínimo de tempo
- Têm avaliação do desempenho medida em termos de transações por segundo (ou milésimo de segundo) e/ou pelo tempo de resposta ao usuário (percentual de transações respondidas em menos de um segundo)
- Suas aplicações são estruturadas com base em duas pré-definições: transações e fluxo de transações. Os trajetos das execuções são calculados e previstos

- Seus esquemas de bancos de dados são muito complexos em termos de número de entidades e de relacionamento entre estas entidades. Os relacionamentos entre entidades impõem ao sistema a dependência múltipla, a integridade referencial e os requerimentos de validação
- Necessitam de edição elaborada no input dos dados a fim de manterem qualidade
- A segurança no acesso é imperativa
- Requerem grande sofisticação no gerenciamento do diálogo
- Utilizam parâmetros ergonômicos para maximizar a produtividade
- Oferecem enorme volumes de aplicações em função do tamanho do banco de dados, número de usuários e de usuários concorrentes e tipos de transações
- Extensivas atualizações e emissão de relatórios fora do horário de pico, dentro de um período de tempo reduzido

A grande vantagem oferecida por este tipo de sistema é a melhoria no desempenho geral dos negócios. Consequentemente, eles possuem, com relativa freqüência, monitoramento e gerenciamento de atividades extraordinárias ou indesejáveis que possam ocorrer em seus subsistemas, com correções posteriores das mesmas.

4.5. Diferenças Fundamentais

O ambiente do sistema operacional difere daquele do Data Warehouse em vários aspectos.

O armazenamento detalhado de assuntos completos em seus dados, requer exatidão, alto nível de estruturação e planejamento no processamento de transações. Ainda, a normalização dos dados afeta a otimização do desempenho que favorece a eficiência e disponibilidade dos dados necessários na execução das atividades diárias da empresa, mas tornam difíceis seu uso no planejamento e na tomada de decisão.Os sistemas operacionais suportados por OLTP mantém os usuários a par das atividades e

das transações básicas da empresa, pois seu propósito principal é produzir respostas à questões rotineiras, manter fluxo de transações e oferecer apoio à análise em um nível organizacional mais elementar. O Data Warehouse, por outro lado, apesar de também se utilizar de dados extraídos dos sistemas operacionais, necessita armazenar e acesso a dados não-voláteis, ou seja, dados estáticos, com histórico de acumulações feitas para séries temporais mais longas e com fins de análise aprofundada. No capítulo 5 serão descritos os principais usuários dos processos on-line: os Data Warehouses. Também serão mostrados seu conceito e o seu modo de armazenamento e gerenciamento de dados.

DATA WAREHOUSE

Os dados tem sido parte permanente das organizações. Embora muitos sirvam meramente como informativos das atividades das empresas em determinados períodos, a grande maioria deles destina-se a servir de apoio à tomada de decisão.

O maior problema com relação aos dados que servem para apoiar decisões sempre relacionou-se a três questões fundamentais: *O que, como e onde armazená-los.*

O advento da tecnologia de Bancos de Dados e de Data Warehouse, esta apoiada em OLAP (On-Line Analytical Processing), trouxe um modo prático, eficiente e rápido a ser aplicado no armazenamento dos dados e usado posteriormente na tomada de decisão para negócios de todos os tipos e tamanhos.

Em tempos não muito remotos o enfoque do recolhimento de dados era dirigido aos sistemas e processos operacionais, cuja arquitetura projetada de arquivo de dados exigia grande performance, diferindo daquela do Data Warehouse, a qual requer uma abordagem estrutural e arquitetura mais voltadas para a flexibilidade e para a larga escala dos dados.

Neste capítulo vão ser apresentados a seleção da arquitetura os tipos de projetos, arquitetura, abordagens de implementação e modelos de armazenamento de dados.

5.1. Armazenamento e Acesso a Dados

Durante a década de setenta, toda tecnologia dirigida para negócios era feita pela IBM em computadores do tipo mainframe e usava ferramentas como Cobol, CICS, IMS e DB2. Os anos seguintes trouxeram as plataformas de mini computadores como o AS/400 e o VAX/VMS. Finalmente, ao fim dos anos 80 o UNIX tornou-se a plataforma mais popular com a introdução da arquitetura cliente/servidor.

A despeito das mudanças processadas nas plataformas, arquiteturas, ferramentas e tecnologias, as aplicações voltadas para negócios continuaram a crescer através do tempo. Hoje, calcula-se que mais de 70% delas ainda se passe em mainframes. A razão principal para tal é que estes sistemas evoluíram a ponto de poder capturar com maior

eficiência e clareza o conhecimento dos negócios das grandes corporações e suas regras, dois fatores importantes e difíceis de retratar até mesmo pelas novas plataformas e aplicações.

Chamados de Legacy Systems, estes sistemas continuam a ser a maior fonte de dados para análise. Os dados armazenados em DB2, IMS, VSAM, etc., para sistemas de transações são direcionados à fitas em bibliotecas remotas de grandes centros de dados. A organização passa, então, a extrair inúmeros relatórios e extratos a cada ano, a partir deles.

Nos anos seguintes, a tecnologia de bancos de dados e sistemas projetados evoluiu até o ponto de se adequar às necessidades de automatização nos negócios. Passou a ser, então descentralizada pela utilização crescente do ambiente cliente/servidor. As bases de dados passaram a ser produzidas para aplicativos do tipo SQL, tais como IBM DB2, Oracle CA/Open-Ingres, Sybase, Microsoft SQL Server e outros produtos do tipo RDBMS. Estes sistemas foram desenvolvidos utilizando uma vasta e variada gama de linguagens ou de ferramentas de desenvolvimento, como aquelas para sistemas de orientação a objetos, com os quais compartilham dados e lógica comuns. Isto possibilitou uma integração cada vez maior da tecnologia de armazenamento e acesso a dados com os requerimentos de orientação para negócios impostos por um ambiente de mercado altamente competitivo.

Atualmente os sistemas funcionam como alavancas para se alcançar o sucesso estratégico das organizações, não apenas eliminando redundância em dados e processos, mas também reduzindo os custos em vários aspectos. Em especial, os de atualização e os de informação consistente, o que era quase impensável em épocas não muito remotas.

5.2. PCs e DESKTOPs

A considerável popularidade aos computadores pessoais (PC – Personal Computer) e desktops nestes últimos anos, incrementou seu uso nos negócios enquanto ajudava a criar novas opções e oportunidades para os mesmos. A explosão de seu uso foi responsável em boa parte pela difusão da tecnologia do Data Warehouse.

Tendo sido inicialmente criados para serem usados como editores de texto e outras tarefas menores, sem qualquer vínculo com funções analíticas, essas máquinas receberam importantes inovações nos últimos anos. Com a ajuda de *softwares* de alta produtividade e com a criação de interfaces gráficas e de aplicações voltadas para negócios, os PCs e desktops tornaram-se o foco principal da tecnologia da computação atual. Incrementadas por hardwares e softwares mais potentes, essas máquinas permitiram o desenvolvimento da arquitetura cliente/servidor (multi-tier architecture), das ferramentas de busca simples e de análise multi-dimensional, além de várias outras ferramentas úteis ao acesso do Data Warehouse.

Os novos processadores trouxeram poder considerável ao mercado da tecnologia de computadores. Sofisticadas arquiteturas de hardware tais como multi-processamento simétrico e chips com maior capacidade de memória foram capazes de potencializar o uso de máquinas consideradas inexpressivas até então. Do ponto de vista da oferta de produtos para os sistemas de Data Warehouse, isto significou preços mais moderados e maior capacidade de armazenamento, pois em épocas não muito distantes armazenar dados significava uma sala cheia de disk drives, o que é impensável em nossos dias quando se pode fazê-lo em discos de apenas um centímetro e meio de tamanho.

Entretanto, surgem dois lados para esta questão: O lado favorável revela que muitas ferramentas de relatório e análise de dados foram re-orientadas para desktops e PCs, possibilitando armazenar e trabalhar com informações extraídas diretamente das fontes de Legacy Systems. O lado menos prático desta situação mostra que este modelo para análise de negócios caracteriza-se por tornar os dados fragmentados e direcionados apenas para necessidades específicas de informações. Ou seja, na falta de padronização que seria necessária às extrações, o usuário obtém apenas a informação requerida em sua busca, o que compromete os requerimentos no caso de haver múltiplos usuários apresentando diferentes necessidades de dados. Além disso, leva-se em consideração a comprovada falta de experiência da maioria dos usuários finais e a grande quantidade de tempo que é necessária para percorrer diversos bancos, tabelas e arquivos, o que ocasiona custos proibitivos.

5.3. Internet, Intranet e Data Warehouse

O fato mais importante depois do intenso desenvolvimento dos computadores pessoais e desktops, foi a explosão da Internet e das aplicações baseadas na Web. Dentre as inovações surgidas pelo uso da Internet, as aplicações de Intranet são aquelas cujo desenvolvimento dentro da indústria da computação tem trazido os mais importantes resultados para Data Warehouse.

Intranets são redes privativas de negócios, projetadas para serem usadas internamente, mesmo sendo baseadas em padrões da Internet. O mix das aplicações Internet/Intranet trouxe importantes influências às aplicações do Data Warehouse. Primeiro porque tornaram os Data Warehouses disponíveis, mundo afora, em redes públicas ou privadas a um custo muito mais baixo do que seria originalmente. Segundo, porque seus padrões de comportamento permitiram que um servidor web ofereça um middle tier onde todo o trabalho pesado de análise se passa antes que o mesmo seja apresentado para ser usado pelo web browser do cliente**.**

A revolução tecnológica causada pelo desenvolvimento maciço de hardwares e softwares, combinado com sua disponibilidade, baixo custo e facilidade de uso tiveram um papel decisivo e de bom impacto para o desenvolvimento no campo do Data Warehouse.

5.4. Conceitos de DW

Para Inmon (1995)o conceito de Data Warehouse sintetiza ao mesmo tempo sua definição e seu papel nas empresas. Segundo o autor, Data Warehouse é um conjunto de dados organizado por tópicos, integrado, que varia com o passar do tempo e não volátil, que serve de suporte para o processo de tomada de decisões da gerência.

Como organizado por tópico entende-se que o Data Warehouse seja orientado para representar as atividades mais importantes da organização tais como produção, vendas, clientes e markenting. As áreas são representadas em dados históricos temporais e de relacionamentos que podem servir como estrutura-chave para a tomada de decisão.

A integração dos dados refere-se a consistência em convenções de denominações, em medidas de variáveis, em estrutura de codificação, em atributos físicos de dados e assim por diante. A denominação usada no Data Warehouse deve ser apresentada em um formato único e integrado e todos os dados que chegarem a sua composição devem ser convertidos ao seu formato.

A característica de variação com o tempo pressupõe que os dados de um Data Warehouse possuam instantâneos ao longo do tempo, ou seja, de períodos que sejam superiores a cinco anos. Isto, segundo Inmon (1995), proporcionaria acuidade aos dados em análise, pois uma vez coletados em horizontes de tempo maiores eles não podem sofrer atualizações (como os dados operacionais), mas mostrará o comportamento dos mesmos em diferentes períodos de tempo.

Os dados utilizados pelo Data Warehouse são normalmente filtrados seja quando são extraídos do ambiente operacional, seja do ambiente externo. Nem todos os dados provenientes de outros ambientes podem ser usados pelo DataWarehouse, mas para que sejam é necessário que tenham características de acúmulo temporal, transformação e resumo, assim terão como principal qualidade a não-volatilidade.

Para Chaudhuri eUmeshwar (1997), Data Warehouse significa uma coleção de tecnologia de apoio a decisões que visa capacitar os trabalhadores do conhecimento (executivos, gerentes e analistas) a tomarem decisões melhores e mais rápidas.

Perkins (1998) apresenta mais um conceito, segundo o qual um Data Warehouse é como um Data Mart, mas usualmente é maior em tamanho e complexidade e aponta seu foco para a empresa como um todo, assim como para as unidades de tomada de decisão. Um Data Warehouse provê dados compreensíveis e de alta integridade sob forma apropriada para suporte à decisões dos usuários finais e formadores de decisão dentro da organização.

Para Gupta (2000), um Data Warehouse é um ambiente extensível estruturado projetado para análise de dado não volátil, logicamente e fisicamente transformado a partir de múltiplas fontes de aplicações para alinhar com a estrutura do negócio, atualizada e mantida por um longo período de tempo, expresso em termos simples do negócio e resumido para análises rápidas.

Finalmente Kimball (1998), apresenta uma sintetizada definição para o termo como uma cópia dos dados de transações, estruturada especificamente para consultas e análise.

A partir dos conceitos apresentados, pode-se concluir que Data Warehouse é, ao mesmo tempo, uma tecnologia e uma poderosa plataforma para combinar dados a partir de antigas e novas aplicações, resolvendo problemas relacionados com os negócios e acionando dados em direção a objetivos estratégicos das organizações.

Diferentemente do mundo dos sistemas operacionais, onde os dados além de serem representados de forma distinta, acomodam múltiplas peças de informação em um só campo e podem ser provenientes de diversas fontes físicas diferentes (lembrando os antigos arquivos de mainframe, as bases de dados não-relacionais, os arquivos indexados e os sistemas baseados em cartões), os dados para Data Warehouse são organizados, em campos de modo a refletir a própria atividade da empresa. Estes dados passam por programas de conversão para serem refinados, editados e reformatados a partir de suas aplicações específicas concernentes ao negócio da organização, sendo, em seguida, armazenados e apresentados como arquivos de Data Warehouse. Este processo visa promover seu alinhamento em torno de áreas de maior importância na organização vindo a ser a chave de sua estrutura como instrumento de consulta para tomada de decisão.

Em função de seu papel dentro das organizações é que os Data Warehouses são projetados para que possam suprir, de maneira atual e otimizada, as necessidades de informação sobre a performance comercial, financeira e operacional das empresas em diferentes momentos.

5.5. Utilização de Data Warehouses

O processo de organizar dados visa não só responder às questões referentes às atividades, como também servir de suporte às decisões concernentes aos negócios da organização. A tecnologia de Data Warehouse suportada por OLAP - On-Line Analytical Processing, é hoje elemento fundamental e estratégico nas corporações.

Como tecnologia da informática, os Data Warehouses ainda são uma tendência de articulação inacabada e em constante evolução. As pesquisas nesta área têm se intensificado à medida que a comercialização de produtos e serviços e o gerenciamento de informações apoiam-se cada vez mais sobre tecnologias de bases de dados.

De acordo com o META Group, Inc., o mercado do Data Warehouse movimentou U$ 8 bilhões de dólares em 1998, entre produtos de hardware, softwares para bases de dados e ferramentas. Sua tecnologia foi usada em vários setores da indústria: Manufatura (pedidos, envios e suporte ao cliente); Varejo (inventário e reposição de estoques); Financeiro (em análise de riscos, de crédito e de fraudes); Transporte (gerenciamento de frota); Telecomunicação (análise de chamadas e fraudes); e Saúde (análise de efeitos).

Atualmente, o setor bancário tem sido líder no uso deste tipo de tecnologia, a qual não só é usada para manter informações sobre o próprio cliente, como também para analisar o seu comportamento.

5.6. Objetivos Estratégicos de DW

As metas fundamentais do Data Warehouse são determinadas pelas necessidades dentro da organização. Em geral elas seguem objetivos estratégicos que são comuns à maioria das empresas. São eles:

- Melhorar a compreensão de problemas
- Diminuir o tempo de espera pela informação requerida
- Incrementar a comunicação
- Reduzir a circulação de papel
- Encolher os custos

- Desenvolver alternativas mais eficientes de repassar informações
- Promover maior controle nos processos de tomada de decisão

As aplicações adicionais do Data Warehouse incluem análise de valor e aproveitamento pelos clientes através do uso de produtos e serviços, melhor gerenciamento de inventários, análise de demanda e de investimentos financeiros pela utilização de dados acurados e de qualidade.

Além destes objetivos, um Data Warehouse deve se mostrar capaz de incrementar serviços aos usuários finais e clientes, com custos reduzidos para os processos dos negócios, aumentando, ao mesmo tempo, sua rentabilidade e flexibilidade para operar e responder mais rapidamente aos desafios competitivos globais.

5.7. Tipos de Projetos para DW

Projetos para Data Warehouse podem oferecer os seguintes benefícios para usuários analíticos:

- Priorizar, na organização, a estruturação adequada dos dados nas consultas analíticas, em vez de tratá-los com simples processamento de transações
- Solucionar e equilibrar as diferenças entre as diversas estruturas de dados em múltiplos e heterogêneos bancos de dados
- Aplicar regras de transformação de dados, validando-os e consolidando-os, quando de sua mobilização de bases do tipo OLTP para OLAP no Data Warehouse
- Manter inalterados os sistemas de produção, quando existirem questões de segurança e desempenho

Os sistemas de Data Warehouses podem ser apresentados em muitas e variadas formas e tamanhos. No presente capítulo serão apresentados os cinco tipos mais conhecidos:

Data Warehouse Virtual - Onde é proporcionado aos usuários finais de terminais ou clientes de workstatations, acesso a bancos de dados operacionais e seus arquivos. Embora esta abordagem permita a capacidade de consultas e geração de relatórios, ela não é recomendável em análises mais complexas de dados quando ocorre a investigação dos negócios da organização.

Data Warehouse Descentralizado ou Departamental - Contém dados informativos, mas de valor apenas para usuários ou grupos específicos. Em geral são chamados de Data Marts e encerram dados capturados de um ou de vários sistemas operacionais. Seus dados são desnormalizados e sumarizados antes de serem propriamente aplicados aos data marts. Neste tipo de Data Warehouse, a abordagem e o processamento dos dados podem ser feitos em sistemas locais, o que incrementa o seu desempenho e disponibilidade. No entanto, à medida que cresce número de data marts, aumentam a redundância, a complexidade de gerenciamento dos dados e do próprio ambiente, sem contar a limitação na flexibilidade dos próprios data marts em satisfazer novos requerimentos de informação.

Data Warehouse Distribuído - Contém várias combinações de data marts em um ambiente simples e distribuído, através dos chamados servidores centrais de middleware. Esta abordagem é limitada por encontrar-se em fase inicial de utilização.

Data Warehouse Central - Inclui dados informativos integrados que são capturados a partir de um ou mais sistemas operacionais, ou até mesmo de

provedores externos de informação. É a abordagem mais comum de Data Warehouse orientada por assunto, cujo objetivo seja a análise dos negócios. É a que apresenta mais facilidade de manipulação do que a aquela que utiliza-se de múltiplos data marts ou do Data Warehouses Distribuído. Os dados contidos neste tipo de DW sempre colhidos a partir de sistemas operacionais, em intervalos definidos pelos usuários para serem detalhados e normalizados posteriormente.

Data Warehouse Two-Tier - É o tipo que emprega tanto um Data Warehouse Centralizado quanto múltiplos data marts descentralizados. É, basicamente, a combinação dos tipos 2 e 4 e a maioria das organizações o empregam são motivadas pelo incremento que seu tipo de arquitetura traz ao Data Warehouse.

Modelos de Data Warehouse devem ter habilidade para organizar informações de negócios a partir de fontes de origem operacional e externas, a fim de torná-las facilmente disponíveis aos usuários. A utilização de um modelo pode auxiliar na definição de uma visão normalizada dos dados existentes, fazendo com que o modelo lógico resultante seja um candidato ao projeto de tabela do Data Warehouse.

5.8. Arquitetura para Data Warehouse

As aplicações de negócios recaem quase sempre ou sobre programas que coletam, criam, modificam, recuperam e/ou apagam dados, ou sobre programas que utilizam, sumarizam, extraem e manipulam dados. Estes dados, quando transformados em informações, podem trazer vantagens competitivas sob forma de um sistema operacional superior, funcionando assim como um excelente objeto de análise para planejamento.

A escolha da arquitetura para o Data Warehouse é, de fato, uma decisão gerencial que irá envolver fatores como a infra-estrutura da organização, seu ambiente de negócios, o gerenciamento desejado, a estrutura de controle, o comprometimento e o

escopo da implementação, a capacidade técnica dos meios empregados pela corporação e os recursos disponíveis. A opção por um determinado tipo de arquitetura é que determinará onde o controle sobre os dados será exercido, se sobre um Data Warehouse central ou sobre os data marts que o comporão e mesmo que a arquitetura venha a sofrer modificações posteriores, é preciso registrar que isso implicará, necessariamente, em maior consumo de recursos e dificuldade para refazer o trabalho de implementação do Data Warehouse. Portanto, selecionar uma arquitetura adequada é também prever o impacto causado sobre variáveis como o tempo de complementar o projeto, o retorno sobre seu investimento, velocidade do benefício de sua realização, grau de satisfação do usuário, potencial de implementação do trabalho, os recursos requeridos em qualquer fase do tempo de estabelecimento do projeto e da própria arquitetura escolhida.

Há três tipos que mais se destacam em termos de arquitetura para Data Warhouse: global, independente e inter-conectada.

A arquitetura global que é aquela capaz de suportar todos ou a maior parte dos dados corporativos integrados, com alto grau de acesso e uso entre departamentos ou linhas de negócios. Por outro lado, o termo global tende a se referir mais ao escopo e ao acesso dos dados do que à sua estrutura física propriamente dita, pois a mesma tanto pode ser distribuída em diversas locações físicas como centralizada em apenas uma delas. Em qualquer dos casos, é bom prever o seu gerenciamento pelo Sistema de Informações da organização, o que não deve significar necessariamente o seu controle por este.

A segunda opção em se tratando de distribuição de dados, pode ser a arquitetura do tipo independente ou stand-alone data marts. Aqui, os data marts são controlados por um departamento ou área de negócios e construídos de acordo apenas com as necessidades específicas destes. Entretanto, pode ocorrer um certo isolamento dos dados, visto que os mesmos são gerados internamente e têm quase sempre a sua orientação voltada para a área a que se destinam. Mesmo assim, muitos deles, em situações bastante comuns e plenamente aceitas no dia a dia das empresas, podem ser procedentes de fontes externas e operacionais.

A terceira opção em termos de arquitetura é chamada inter-conectada. Nela, a implementação é que é distribuída e embora os data marts sejam separados por áreas, eles podem ser integrados e conectados a fim de prover sua melhor visualização. O que acontece de fato é que os dados inter-conectados podem vir a formar uma espécie de Data Warehouse global, onde os usuário de um departamento podem acessar facilmente os dados dos data marts de outros departamentos. Mas é preciso ressaltar que esta estrutura, além da tendência para tornar-se mais complexa, pode estimular a redundância dos dados, sendo preciso criar uma espécie de grupo de gerenciamento e controle daqueles que devem se tornar dados comuns a múltiplos ambientes.

Em alguns casos e, de acordo com as necessidades da organização, é possível combinar dois diferentes tipos de arquitetura. Este procedimento tem se tornado popular, pois, por definição, torna menos rígida a estrutura do Data Warehouse, além de se poder tirar proveito das diversas vantagens de cada um dos tipos escolhidos para o projeto.

5.8.1. Elementos da Arquitetura de DW

Os Data Warehouses foram os pioneiros no armazenamento de informações gerenciadas através de sistemas de bancos de dados e da computação em redes distribuídas. Através deles, é possível acessar bancos multi-aplicativos, onde as informações são disponibilizadas e os custos de informação se reduzem significativamente.

Com respeito à arquitetura do Data Warehouse, os conjuntos de elementos que a compõem variam nas opiniões dos mais conhecidos pesquisadores.

Para Tanler (1998) o Data Warehouse é composto dos seguintes elementos em sua infra-estrutura:

1. **Armazenamento de Dados** – Que equivalente a um depósito estruturado de dados, onde é gerenciado, especificamente, o seu conteúdo.
2. **OLAP (On-line Analyse Processing)** – Refere-se ao conjunto de ferramentas on-line necessárias ao acesso e análise de dados. É onde se

destacam as funções de pesquisa e relato de resultados, análise (em geral, multi-dimensional) e mineração de dados (Data Mining).

3. **Tecnologias de Internet** – Trata-se, mais especificamente da tecnologia de Intranets, onde são evidenciados o melhoramento e a importância da comunicação e da colaboração dentro entre usuários dentro da empresa.

Kimball et al. (1998), sugerem uma arquitetura para Data Warehouse mais voltada para onze elementos fundamentais:

1. **Sistema de Fontes (Source System)** – Sistema operacional de registros cuja função é capturar as transações nos negócios. É comumente chamado de Legacy System, em ambientes do tipo mainframe e suas prioridades são atualização e disponibilidade de dados. As consultas são reduzidas e restritas tanto em termos de fluxo quanto em termos da própria demanda sobre sistema.

2. **Plataforma de Dados (Data Staging Area)** – Conhecida como sendo a área e o conjunto de processos que refinam, transformam, combinam, unificam (não permitindo duplicação), arquivam e preparam a fonte de dados para uso do Data Warehouse. Esta plataforma é o item de ligação entre o sistema de fontes e a apresentação dos dados em si. É dividida em um certo número de máquinas e nem sempre é orientada pela tecnologia relacional.

3. **Servidor de Apresentação** – É no servidor de apresentação que os dados são armazenados e apresentados. No caso dos servidores baseados em bancos de dados relacionais, as tabelas são organizadas em esquemas do tipo Star. De outro modo, para servidores baseados em tecnologia MOLAP, os dados são organizados em modelos dimensionais.

4. **Data Mart** – É o subconjunto lógico, baseado em dados granulares, que complementa o Data Warehouse. Normalmente o data mart é construído em torno

de uma parte específica do negócio da empresa e é organizado em torno do mesmo. Podem existir muitos data marts e o conjunto deles é que irá compor uma das partes mais importantes do Data Warehouse. A experiência dos autores tem mostrado que o modelo dimensional é o mais adequado e robusto no que diz respeito à construção dos data marts na arquitetura de apoio ao Data Warehouse.

5. **Armazenador de Dados Operacional (Operational Data Store)** – Definido por diversos autores como sendo uma espécie de depósito para os mais diversificados e incompatíveis requisitos para negócios, que incluem desde o armazenamento de dados voláteis, passando por seu refinamento até aonde o dado, cujo nível, por mais básico que se apresente, seja objeto de uma transação.

6. **ROLAP (Relational OLAP**) – Conjunto de interfaces e aplicações que oferecem textura dimensional aos bancos de dados relacionais.

7. **MOLAP (Multidimensional OLAP)** – Conjunto de interfaces, aplicativos e tecnologias proprietárias de bancos de dados para usuários, com textura múltipla em termos de dimensão. Pode ser entendido como sendo um conjunto de ferramentas que fazem consultas, analisam e apresentam informações, cujo alvo é dar suporte às necessidades dos negócios da empresa. O mínimo a ser oferecido em se tratando de MOLAP é o conjugado de pacotes gráficos, ferramentas de acesso a dados e de editores de relatórios, além, é claro, de sua facilidade e descomplicação em termos de apresentação na tela dos microcomputadores.

8. **Ferramentas de Acesso Para Usuários Finais** – Clientes de Data Warehouse relacionais mantém sessões de contato constantes com os servidores de apresentação, o que torna imprescindível, por exemplo o uso de SQL na exposição dos dados. Eventualmente, após a consulta, os dados podem ser exibidos sob forma de relatório, gráficos, formulários de análise, ou de maneira mais complexa como a tecnologia de data mining.

9. **Ferramentas de Consulta Ad Hoc** – Tipo específico de ferramenta de acesso a dados que estimula o usuário a formular seus próprios requerimentos por informações, pela manipulação direta das tabelas relacionais e suas junções. Embora ofereçam flexibilidade e incremento em seu uso, comprovou-se que este tipo de ferramenta pode ser usada efetivamente por apenas dez por cento de todos os seus usuários potenciais.

10. **Aplicações de Modelagem** – Tipo sofisticado de arquitetura cliente para Data Warehouse que possui capacidade analítica para transformar ou assimilar de forma sumarizada os dado de saída. Seus modelos de aplicação incluem: *Previsão; Pontuação por Comportamento; Alocação de Modelos*; *e Data Mining.*

11. **Metadado** – Todas as informações concernentes ao ambiente do Data Warehouse, que não sejam os dados em si mesmos. Korzybski (1996), traduz metadado como sendo o instrumental que transforma dado bruto em conhecimento. Portanto, metadado é o campo que determina o fluxo e que alinha os dados, de modo que venham a fazer sentido dentro do Data Warehouse.

White (1995), ainda nos propõe alguns elementos simples na composição arquitetural do Data Warehouse:

1. **Componente de Desenvolvimento** – Tem a finalidade de projetar as bases de dados usadas no Data Warehouse e os aplicativos de captura de dados provenientes de fontes operacionais e/ou externas.

2. **Componente de Aquisição de Dados** – Objetivam capturar dados a partir de fontes ou coleções de dados, refiná-los, transportá-los e aplicá-los às bases de dados do Data Warehouse.

3. **Componente de Gerencimento de Dados** – Administram todas as operações relativas ao Data Warehouse.

4. **Componente de Distribuição de Dados** – Operam como distribuidores de dados do Data Warehouse para data marts e sistemas externos à organização.

5. **Componente de Informação de Diretório** – Tem por finalidade prover informação sobre o conteúdo e significado dos dados contidos no Data Warehouse, para os usuários em geral.

6. **Componente de Acesso a Dados** – Componentes que provêem os usuários finais com as ferramentas necessárias para que possam acessar e analisar os dados do Data Warehouse.

A figura da página seguinte mostra o relacionamento entre os componentes da arquitetura do Data Warehouse.

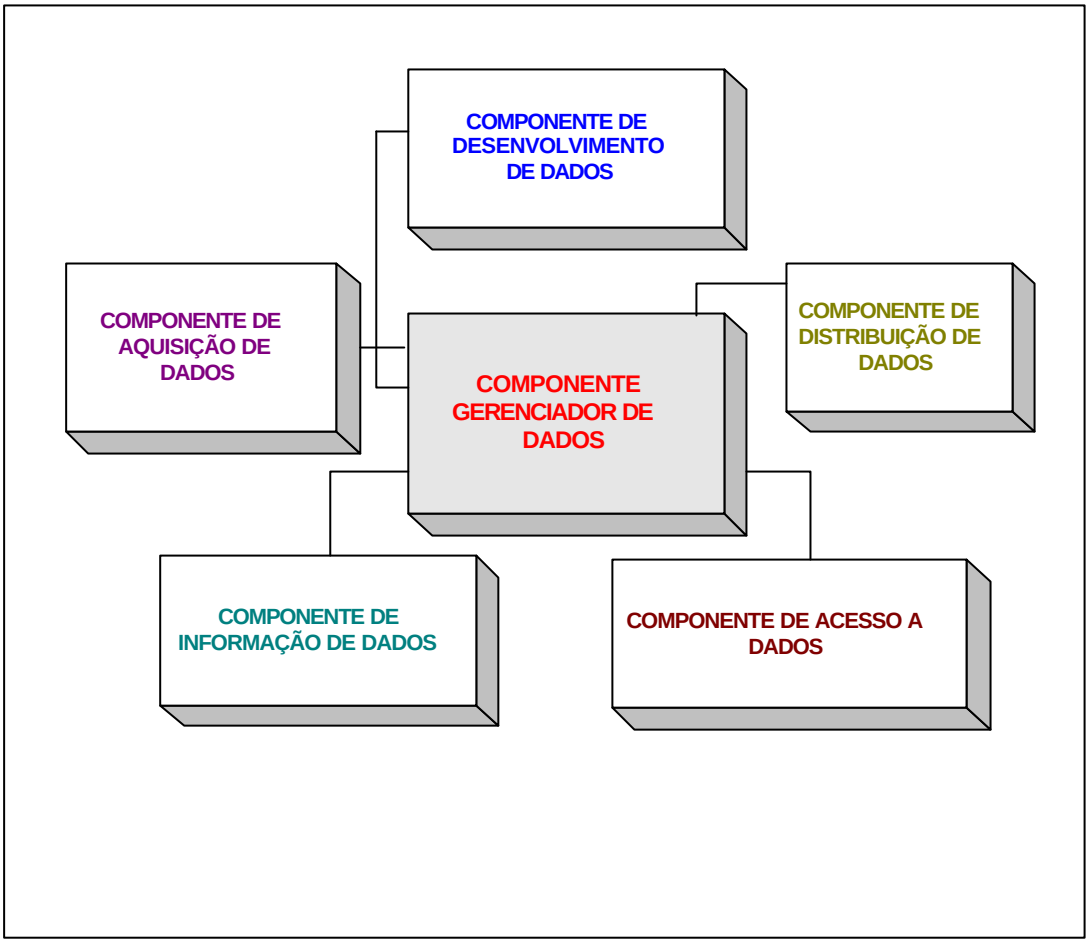

Figura 5.8.1.1. Componentes da Arquitetura de Data Warehouse

De um modo geral, os elementos fundamentais na arquitetura do Data Warehouse variam tanto em termos da própria existência, quanto à nomenclatura que lhes é aplicada. Entretanto, os elementos essenciais são aqueles que, em primeiro lugar, coletam, refinam e gerenciam os dados. Segundo, é imperativa a existência de componentes, ou elementos que forneçam acesso on-line a esses dados. E por último, é fundamental que exista um meio de apresentação sob forma de facilitar a análise, quando houver necessidade de tomarem-se decisões.

A maior parte dos produtos colocados no mercado não possui todos os elementos aqui descritos e nem oferece uma solução completa de produtos para a operacionalização eficiente de um Data Warehouse. Entretanto, muitos dos desenvolvedores de produtos

para DW procura projetá-los de modo que possam se integrar com os de outros fabricantes, oferecendo, deste modo soluções mais completas para as organizações que se utilizam do armazenamento de dados para fins de análise.

5.9. Abordagens de Implementação para DW

As abordagens de implementação podem ser de dois tipos: Top Down ou Bottom Up. No primeiro caso, a implementação requer maior planejamento e adequação sobre o projeto nas fases iniciais. O envolvimento do pessoal de diversas áreas da empresa é imperativo para decisões concernentes à utilização de fontes de dados, segurança, estrutura, qualidade, padronização e modelo de dados que, tradicionalmente, devem estar prontos antes que a fase de implementação tenha começado.

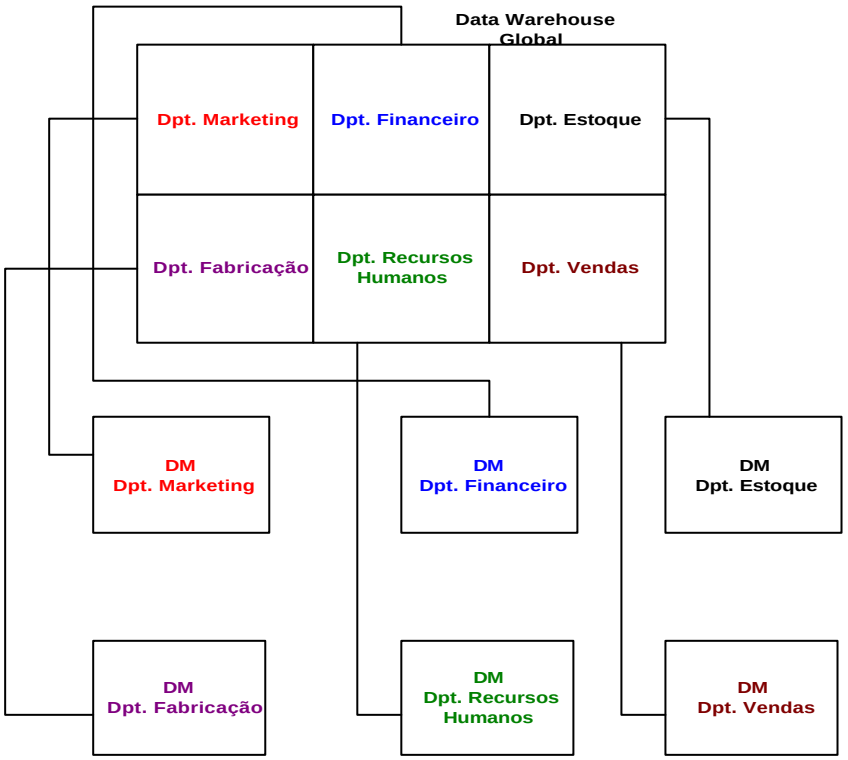

Figura 5.9.1.1. Data Warehouse Global e Data Marts na Abordagem Top Down

A abordagem top down é reconhecidamente melhor utilizada e preenchida pela arquitetura global de distribuição de dados. Uma de suas principais características é que ela resulta em dados mais consistentes e completos, pois as regra de negócios são claramente definidas desde que se dá início ao projeto. Por outro lado, uma considerável desvantagem reside no fato de que seu custo é alto em função do tempo de planejamento requerido, principalmente para que as diversas áreas da organização estejam de acordo quanto a definição dos dados e regras gerenciais a serem utilizados.

Ultimamente, a abordagem top down vem caindo em desuso em função da necessidade de utilização de um Sistema de Informações centralizador que seja responsável por todos os recursos de hardware, o que dificulta e torna maiores seus custo, além de aumentar o seu tempo de implementação.

A segunda abordagem ou implementação bottom up, envolve planejamento e projeto de data marts que não necessitam que uma infra-estrutura global seja erguida anteriormente. Isto quer dizer que os data marts podem ser construídos antes ou paralelamente ao Data Warehouse global.

Os dados para este tipo de abordagem são retirados tanto do DW global, quanto de sistemas operacionais ou fontes externas de dados. Por isso a sua implementação, que começa com um data mart, pode se expandir através do tempo e seus resultados, até agora, têm se mostrado muito mais satisfatórios que aqueles da abordagem top down.

Embora a abordagem bottom up apresente muitas vantagens, é preciso considerar que fatores comuns a sua estrutura, tais como a redundância e a inconsistência de dados podem comprometer a sua eficiência, sendo, portanto, imperativo que haja maiores cuidados em seu planejamento, monitoramento e no estabelecimento de diretrizes que poderão promover o seu melhor desempenho.

Atualmente, a abordagem bottom up tornou-se a melhor escolha das organizações em função de seu retorno econômico sobre hardware, bem como sobre o tempo de espera dos usuários pela sua implementação.

A seguir é mostrada a figura da abordagem bottom up com seus Data Warehouse e data marts.

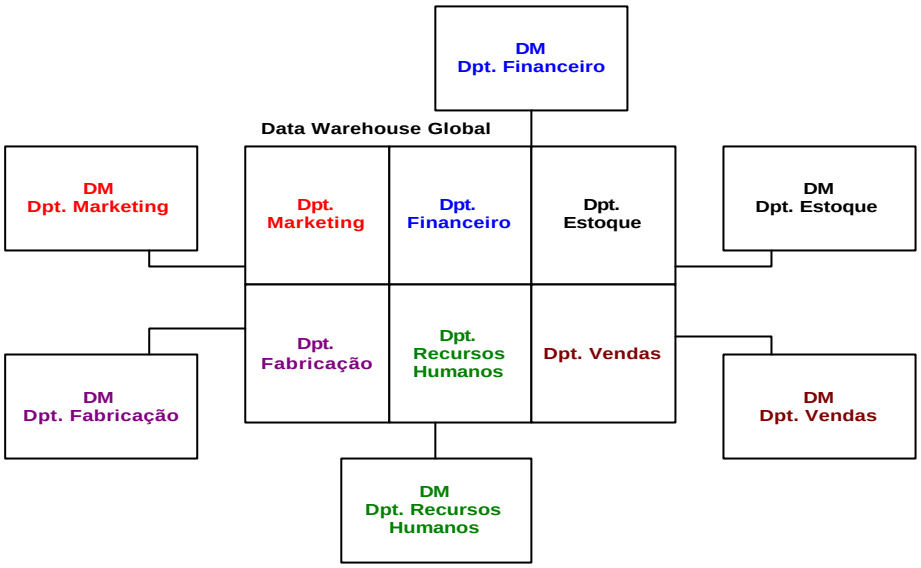

Figura 5.9.1.2. Data Warehouse Global e Data Marts na Abordagem Bottom Up

Finalmente, depois de vistos tanto os aspectos positivos quanto os negativos de cada abordagem, pode-se considerar como uma boa opção a combinação de ambas, o que, embora mostre-se um difícil ato de equilíbrio, pode também se apresentar de forma gerenciável com adequados planejamento e projeto. Para que isto ocorra, é preciso desenvolver o nível básico na definição da infra-estrutura do Data Warehouse. Por exemplo, o primeiro passo seria identificar as linhas de negócios que estariam participando. Com esta informação, seria possível tornar mais clara a visualização do negócio da organização, seus processos e dados por áreas de interesse, assim como a identificação dos elementos necessários para a implementação dos data marts. O passo a ser seguido paralelamente à implementação dos múltiplos data marts, é o desenvolvimento de um plano para estabelecer as regras de utilização de seus dados. Isto pode significar tanto a criação de uma estrutura global de Data Warehouse quanto simplesmente o armazenamento comum e mais acessível dos dados para todos os data marts envolvidos no projeto. Ressalta-se que, em alguns casos seria apropriado duplicar os dados a fim de incrementar sua disponibilidade. Mas esta é uma decisão que deve levar em consideração aspectos como o espaço requerido para o armazenamento, a facilidade de acesso e o impacto da redundância em relação aos requerimentos de

manutenção de dados em diversos data marts quando se visa sobretudo a conservação da consistência dos mesmos.

A escolha da combinação das abordagens *top down* e *bottom up* é uma saída que pode resolver muitos problemas de implementação de um Data Warehouse. Como mencionado anteriormente, um cuidadoso planejamento e o constante monitoramento são fatores importantes que concorrem para um melhor resultado.

5.10. Dados para Data Warehouse

Os dados essenciais ao funcionamento do Data Warehouse se referem aos negócios e podem ser extraídos de fontes internas, como os sistemas operacionais da empresa, e de fontes externas quaisquer que se relacionem com as atividades da organização.

Um infinito número de empresas, incluindo as de atacado, varejo, finanças, seguros e saúde, mantém rotineiramente enormes volumes de dados sobre suas atividades e clientes. Implícitos nestes dados estão modelos típicos de comportamento destes clientes, conhecidos por auxiliarem as organizações na orientação de suas estratégias de marketing em reduzir riscos financeiros e incrementar o campo de ações através de processos de tomada de decisão.

A qualidade das decisões a serem tomadas na empresa depende diretamente da qualidade dos dados que apoiam seu sistema de informações. Normalmente, as aplicações destes dados relacionam-se às atividades de coleta, análise, relatório e informação compartilhada e orientada e contém onze atributos mais importantes:

- São provenientes de bancos de dados estáticos
- Possuem longos históricos de acumulação (consistência temporal)
- Têm consistência global
- São resumidos
- Passam por restaurações periódicas, visando evitar perdas importantes de informações
- Têm formatação e recuperação simplificada

- Podem ser importado e exportados de e para outras fontes
- São armazenados em seqüências de tempo explícitas
- Não permitem alterações instantâneas
- Possuem magnitude superior à da sua fonte de extração
- São compartilhados entre muitos usuários.

Nas últimas três décadas os dados do DW têm sido armazenados eletronicamente. As principais razões para tal comportamento são: a facilidade de acesso aos dados e o crescimento exponencial (em gigabytes e até mesmo terabytes) dos mesmos em termos de armazenamento.

De fato, os dados no ambiente do Data Warehouse possuem múltiplos propósitos e usuários, todos com necessidades variadas em relação ao seu uso. Os dados que suportam as atividades do processo decisório devem ser significativos e corretamente armazenados. Entretanto, se um conjunto particular de dados pode servir unicamente a um nível mais elementar de atividade da organização, tal como as atualizações constantes no fluxo de caixa, por exemplo, outros seguem prioridades mais elevadas, tal como o registro histórico de operações financeiras de alto risco e longo prazo realizadas pela organização.

Aqui, a questão sobre a escolha dos dados depende das funções a que se destinam e para que haja lógica e maximização na eficiência, tanto na escolha, quanto no armazenamento, ou no uso dos mesmos, é preciso que se faça uma análise mais profunda do modelo de armazenamento de dados para Data Warehouse, a ser usado pela corporação.

5.10.1. Características de Dados para DW

A grande variedade de dados relacionada às funções da organização provém de uma enorme mistura de fontes operacionais tais como registros de clientes e informações de produtos, entre outras coisas. Misturadas, estas informações perdem a relevância necessária a relatórios de negócios, ao mesmo tempo que se torna difícil a sua consulta.

A solução, então, é a orientação dos dados por assunto de modo que, organizados estes, sua execução venha ser produtiva.

Data Warehouse é o elemento que consolida dados operacionais de uma grande variedade de fontes, com nomeação convencionada, medidas, atributos físicos e semântica sólidas. Isto, porque a estabilidade desses dados é que vai permitir que toda a organização, partindo de que fontes forem, tenha referências cruzadas eficientes e forneça aos analistas a melhor compreensão possível dos negócios. Estas características dos dados são chamadas de consolidação e consistência. Outra importante característica dos dados para DW diz respeito à disponibilidade ou fato dos mesmos poderem ser instantaneamente carregados e consultados, ou seja, que sejam oferecidos prontamente ao usuário em função analítica e não apenas em funções de exclusão, inserção e atualização. Este é, dentre todos os processos, o que melhor define o carregamento dos dados ou data load.

As características acima mencionadas só se tornam realidade para o Data Warehouse quando se parte para uma classificação mais objetiva dos dados a serem armazenados em sua estrutura. Assim, são categorizados a seguir três tipos básicos de dados que podem ser usados para satisfazer os requerimentos da organização. São eles:

- **Dados de Tempo Real ou Real-Time Data** - Representam o status atual do negócio. São bastante utilizados em aplicações operacionais que sustentam o dia a dia da empresa e estão em constante modificação enquanto estas se processam. Possuem alto nível de granularidade em virtude de seu detalhamento e são normalmente acessados nos modos read/write pelas transações operacionais. Seu uso em Data Warehouses é condicionado ao refinamento que proverá a sua qualidade como peça de informação. Mesmo assim, como podem ser provenientes de múltiplas fontes, eles podem não apresentar a consistência e o significado necessários para a função de análise. Por exemplo, no sistema bancário taxas de câmbio de

moedas podem apresentar diferentes medidas entre sistemas, obrigando a correção destas anomalias para perfeita compreensão da informação.

- **Dados Derivados ou Derived Data** - É o dado que é criado a partir de algum processo tal como a sumarização, média ou agregação dos dados do tipo real time. Estes dados possuem o nível mais baixo de granularidade exigido por um Data Warehouse e representam instantâneos de momentos específicos no tempo ou de fatos que sustentam uma situação real para a organização. Seu acúmulo vai de períodos de três a cinco anos, renováveis com tanta freqüência quanto necessária para dar suporte aos papéis informativo e analítico que raramente necessitam de grandes volumes de dados detalhados.

- **Reconciled Data** - É o conjunto de dados de tempo real que passou pelos procedimentos de limpeza, ajustamento e acréscimo para ser capaz de se tornar uma fonte qualificada de dados a ser usada pelos analistas. Seu requisito básico de qualidade é sempre a consistência. Na verdade, este tipo de dado é raramente definido explicitamente, mas pode-se afirmar que ele resulta em um tipo especial de dados derivados ou um resultado lógico de operações derivativas. Algumas vezes, estes dados são armazenados apenas como arquivos temporários requisitados na transformação consistente de dados operacionais.

Muitas metodologias podem ser usadas na estruturação dos dados para DW, mas a utilização de um modelo de dados voltado para a definição de todos os elementos

comuns aos negócios, com uma visão de alto nível do seu gerenciamento pode ser um fator que promove o entendimento dos requerimentos da articulação das informações para Data Warehouse.

5.10.2. Metadados

Outro fator de grande contribuição para a compreensão dos negócios da organização no ambiente do Data Warehouse é a composição de metadados. Metadados, ou dados sobre dados, são, por assim dizer, o sistema nervoso central do Data Warehouse. Através deles pode-se exercer gerenciamento e controle na criação do DW, embora normalmente eles estejam situados fora deste último. Para os usuário, metadados são como uma espécie de catálogo dos assuntos no Data Warehouse, havendo dois tipos de metadados, a saber:

- **Metadados Estruturais** - Usados na criação e manutenção do Data Warehouse, são capazes de contar em detalhes a sua estrutura e conteúdo. Descrevem as entidades de dados, suas características e o modo como estes estão relacionados uns aos outros. Promovem uma visão da melhor utilização dos dados, do significado de sua documentação e dos planejamentos estratégico e operacional dos dados da organização. Metadados estruturais também incluem o desempenho métrico dos programas e buscas para que os usuários e desenvolvedores possam saber quais seus tempos de espera.

- **Metadados de Acesso** - São os links dinâmicos entre o Data Warehouse e as aplicações para usuários finais. Em geral, contém um dicionário de termos padrões e as medidas organizacionais suportadas pelo DW. Incluem também as descrições e locações dos servidores, bancos de dados,

> tabelas, dados detalhados e fontes de dados originais. São capazes de prover regras para atividades drill up e drill down, das visões multidimensionais e da hierarquia dos assuntos tais como produtos, mercados e consumidores. Dados de acesso também proporcionam regras de cálculos para os usuário que são clientes, bem como para suas buscas. Neles estão contidos fatores como segurança para a visualização dos trabalhos individuais, em grupos, assim como permissões para alterações, distribuições, sumários e outras análises.

Metadados, por definição, são os dados que ajudarão os usuários do Data Warehouse a encontrarem as informações requeridas pela função de análise dos processos decisórios. Eles apoiam a busca pelos dados enquanto oferecem instruções mais detalhadas dos dados e orientam seus consumidores a tomarem o melhor caminho entre todos os caminhos oferecidos pelo DW.

5.10.3. Armazenamento de Dados

Para tornar-se útil, um Data Warehouse deve sintetizar toda a dinâmica de dados que lhe ocorre, ser compreensível, além de capacitado a prover um retrato completo e abrangente na estruturação dos mesmos. Para suprir tantas exigências é preciso projetar o modelo onde estejam contidos tanto os representantes físicos quanto as coleções de dados derivados, sumarizados em pacotes agregados e produzidos para satisfazer as necessidades de acesso dos usuários. Depois de tudo, o modelo também deve incluir arquivos de dados e metadados para qualquer que seja sua extensão histórica manejável.

Algumas opções de múltipla armazenagem de dados são:

- **ROLAP** – O armazenamento é feito sob forma de um cubo e contém grandes volumes de dados básicos e detalhados. Além disso, desde que ROLAP trabalha com data marts relacionais, a

escalabilidade deste modelo é limitada apenas pelos procedimentos de armazenagem relacional, o que permite maximizar seu investimento em tecnologia relacional e ferramentas de gerenciamento de dados corporativos. Entretanto, seus usuários devem esperar por uma certa inconsistência nos dados, como também tempos de resposta relativamente lentos.

O projeto inicial do modelo ROLAP é baseado em técnicas dirigidas de bancos de dados, os quais seguem os seguintes passos:

Construção de modelo utilizando técnicas como desnormalização e introdução de esquemas do tipo Star e Snowflake;

Adição de dados com sumarização e agregação apropriadas;

Divisão de grandes conjuntos de dados em porções menores e manuseáveis a fim de incrementar o desempenho;

Introdução de índices bitmap visando aumentar a performance, mesmo que resulte em alargamento do tamanho do banco de dados e do tempo que se levará para construir os índices;

Criação e armazenamento de metadados, que incluem definição das dimensões, seu mapeamento às tabelas relacionais, descrição das hierarquias e seus relacionamentos, definição das funções de sumarização e agregação, fórmulas e cálculos, monitoramento, etc.

- **MOLAP** – Os dados básicos de um cubo são armazenados com os dados de agregação em uma estrutura multidimensional de alto desempenho. O armazenamento MOLAP fornece excelentes desempenho e compactação de dados, pois a visão multidimensional, que raramente usa menos do que três

dimensões, provê o modelo com a habilidade de "fatiar" as referidas dimensões, enquanto fundamenta o processo analítico em flexibilidade de acesso às informações. As atividades iniciais de projeto e estabelecimentos são orientadas um plano lógico ou a um modelo de informação. Seus passos básicos são:

1. Seleção de uma das funções do negócio, tal como análise de vendas ou relatórios financeiros;
2. Identificação de valores numéricos a serem armazenados, como renda sobre as vendas e vendas por clientes;
3. Determinação das dimensões (tempo, produto, cliente, etc.) e a granularidade de cada uma delas. Por exemplo, tempo por mês, trimestre e produtos, por família de produtos ou classes;
4. Definição de um modelo lógico e do carregamento dos dados multidimensionais armazenados, seja diretamente da fonte, seja através da filtragem ou combinação dos conteúdos selecionados do Data Warehouse ou dos data marts.

- **HOLAP** – São utilizados dois modelos de armazenamento de dados para compor HOLAP. Enquanto os dados básicos de um cubo são depositados em um banco de dados relacional, os dados de agregação permanecem sob forma de estrutura multidimensional de alto desempenho. As opções adicionais do HOLAP incluem cubos virtuais e partições. O armazenamento HOLAP oferece os benefícios do MOLAP para agregações, sem necessidade de duplicação dos dados detalhados básicos.

Dos modelos apresentados, o modelo híbrido é aquele que melhor fundamenta o Data Warehouse, pois sua criação deve-se ao uso de mais de uma metodologia. De fato, os melhores modelos são aqueles que sintetizam diferentes técnicas e onde cada um

contribui com uma parte para o resultado como um todo. No modelo híbrido é representada a diversidade de diferentes repositórios de dados requeridos na aquisição, armazenamento, empacotamento, entrega e compartilhamento de informações. Os outros requisitos para o modelo de Data Warehouse relacionam-se às representações físicas de dados e à existência de metadados que não serão vistos aqui.

5.10.4. Arquitetura de Dados

A coleção de dados correntes orientados e integrados que compõem o Data Warehouse, além de serem considerados sua espinha dorsal, são os elementos que farão frente ao processo decisório da empresa. É então, compreensível a importância da arquitetura dos mesmos para as funções de análise.

Em relatório recente, o META Group observou que a má qualidade dos dados é responsável por 41% dos fracassos de Data Warehouses. Normalmente, os dados que se prestam como peças de informação vêm diretamente de fontes como os sistemas operacionais ou fontes externas e podem ser agregados como dados brutos ou organizados por áreas representando o todo da organização, para serem, em seguida, disponibilizados com finalidade específica de manter informados os seus usuários. Conclui-se, pois, que o maior desafio dos grupos de analistas envolvidos com a articulação do Data Warehouse seja a triagem dos dados que irão compor este último. Conceitos básicos como capturar, limpar, refinar, sumarizar, agregar dados e posicioná-los em uma estrutura ótima de acesso são fundamentais para tarefa de informar que vão possuir estes dados. Assim, é necessário planejar a arquitetura dos dados contidos no modelo de Data Warehouse.

A Teoria da Engenharia de Dados, que é a disciplina que estuda como modelar, analisar e projetar dados com o máximo de utilidade indica que há quatro ambientes genéricos de dados:

- **Arquitetura de Dados Dedicados:** Cada aplicação tem um conjunto de arquivos projetado particularmente. A estrutura dos

dados é firmemente encaixada à aplicação da qual os arquivos de dados são propriedade.

- **Arquitetura Fechada de Banco de Dados:** O Sistema de Gerenciamento de Banco de Dados (SGBD) é usado como forma de prover vantagem tecnológica – visão, segurança, atomicidade, recuperação, chaveamento, etc. - sobre o sistema de arquivos, embora sejam utilizadas bases de dados independentes, distintas e separadas para cada aplicação. O gerenciamento do sistema é empregado como uma espécie de propriedade privativa e poderosa da aplicação gerenciada pelo SGBD. Entretanto, existe um alto grau de redundância de dados, bem como uma certa ineficiência na administração dos mesmos. A interface com cruzamento e movimentação de dados entre bases fechadas de dados tem que, ocasionalmente, converter, editar e/ou reestruturar dados enquanto os movimenta entre as definições proprietárias.

- **Arquitetura de Dados Por Assunto:** Os dados, baseados em seus atributos internos, são analisados, modelados, estruturados e armazenados, independente de qualquer aplicação específica. Eles também constituem em uma fonte compartilhada através da função de administração que é sua proprietária, para todos os usuários potenciais. Nas operações do dia a dia, este tipo de arquitetura tem extensiva utilização e seu compartilhamento ocorre através de visões sensíveis de suas aplicações.

- **Arquitetura de Suporte à Decisões***:* Os bancos de dados são construídos para buscas rápidas, recuperação, pesquisas *ad hoc* e facilidade de uso. Os dados são, normalmente, extrações periódicas de bases de dados por assunto Para minimizar o número de

> extrações e incrementar a consistência da relação tempo/conteúdo, os dados são compartilhados em níveis corporativo, departamental e local. As definições dos dados são mantidas sincronizadas com a fontes de bases de dados, visando assegurar a habilidade de inter-relacionar dados com as extrações provindas de bancos de dados múltiplos e orientados à matérias, sem necessidade de refiná-las. Bancos de Dados de Suporte à Decisões são usados para analisar a empresa.

O planejamento e Arquitetura do Data Warehouse o pressupõem o objeto orientado para o plano estratégico dos negócios da empresa e não apenas a utilização da tecnologia para acesso à informação. Isto significa dizer que, é preciso antecipar a necessidade de dados baseados no mercado e sua potencial utilização, assim como incluir todo e qualquer tipo de dado de reconhecido valor, procurando identificar seus consumidores, particularmente aqueles com difícil acesso às fontes, e deste modo, concentrar os esforços na sua disponibilidade e flexibilidade enquanto se gerencia performance e volume.

5.11. TIPOS DE MODELAGEM DE DW

A maior parte das tarefas que envolvem dados relacionam-se a questão de como os bancos de dados podem ser projetados de modo a suportar as necessidades dos usuários do Data Warehouse. Para suprir esta carência, foram sendo propostas várias técnicas ao longo dos anos e entre as várias existentes, há duas técnicas que serão discutidas neste trabalho: a modelagem relacional e a modelagem dimensional. A primeira será apresentada apenas a título informativo, já que aquela que de fato vai ser de maior interesse na proposta de um modelo de Data Warehouse, será a modelagem dimensional.

Um modelo é, em geral, uma abstração e um reflexo do mundo real. A modelagem proporciona a habilidade de visualizar o que ainda não foi concretizado. Tradicionalmente, os modeladores, como podem ser chamados, fazem uso de diagramas

como técnicas de modelagem, a fim de poder comunicar os negócios aos usuários analistas. É o que acontece com os dados do modelo relacional, cujo diagrama é uma ferramenta que coopera tanto com os requerimentos de análise quanto no projeto resultante da estrutura dos dados.

Se for considerado que a bem organizada modelagem de dados é uma abstração satisfatória da informação contida no Data Warehouse, então é natural que um modelo seja o melhor método para entender e gerenciar os negócios da organização. Através das técnicas de modelagem é possível se criar um plano, conjuntos de padrões e linhas de conduta para se implementar o Data Warehouse.

A seguir serão revistas brevemente as duas modelagens de armazenamento de dados mais utilizadas pelos projetistas de Data Warehouses: a modelagem relacional (MR) e a modelagem dimensional (MD).

5.11.1. MODELAGEM RELACIONAL

Conhecido como um modelo de dados para aplicativos comerciais, o modelo relacional procurou, em primeiro lugar, dar enfoque à eliminação da redundância de dados e à manutenção da consistência entre diferentes fontes de dados e aplicações. Este modelo é representado por um diagrama que se utiliza de três símbolos gráficos básicos: entidade, relacionamento e atributo, que são assim definidos:

- **Entidade** refere-se a uma pessoa, lugar, coisa ou evento de interesse dos negócios ou da organização. A entidade pode representar uma classe de objetos, que sendo coisas pertencentes ao mundo real, podem ser observadas e classificadas por suas propriedades e características. Exemplos de entidades são produtos, modelos de produtos e componentes de produtos.

- **Relacionamento** é descrito como a interação estrutural e a associação entre as diversas entidades de um modelo. A relação entre duas entidades pode ser definida em termos de cardinalidade,

ou seja, o máximo número de instâncias de uma entidade que está relacionada a uma instância em outra tabela ou vice-versa. Exemplos possíveis de cardinalidade são: relacionamentos de um-para-um (1:1), um-para-muitos (1:M) ou muitos-para-muitos (M:M).

- **Atributos** são descrições das características de propriedade das entidades. É importante que os atributos usem nomes convencionados e auto-explanatórios. Isto quer dizer que um nome de atributo deve ser único, de modo a evitar que sejam confundidos as características de diferentes entidades.

Outro importante conceito do modelo relacional é a normalização. Normalizar é o processo de conceder atributos à entidades, de modo que a redundância seja reduzida ao mínimo, evitando-se anomalias, provendo-se uma sólida arquitetura na atualização dos dados e reforçando-se a integridade do modelo de dados. Um bom exemplo de normalização é o processo de determinação dos relacionamentos do tipo muitos-para-muitos (M:M).

A modelagem relacional se utiliza de diagramas para demonstrar várias entidades distintas que são transformadas em tabelas. Os registros de cada tabela são feitos em separado, transformando-se em um novo diagrama mais complexo. Com freqüência, não se pode saber o número exato de tabelas, seus relacionamentos, sua importância, ou os valores numéricos que armazena. Por esta razão e mais algumas características como a simetria das tabelas e a finalidade maior de oferecer suporte ao processamento de grande quantidades de transações, este modelo é reconhecidamente mais utilizado para transações OLTP e não possuem tanta relevância para o ambiente do Data Warehouse quanto o modelo dimensional nas transações do tipo OLAP. Portanto, as demais características e técnicas da modelagem relacional foram descritas sumariamente neste capítulo, visto que a pouca importância de sua compreensão no presente trabalho.

- **Medida** é o atributo numérico de um fato que representa a performance ou comportamento do negócio relativo àquela dimensão. Na verdade, são chamados de varáveis e determinadas pela combinação dos membros das dimensões e estabelecidas nos fatos das tabelas. Exemplos: moedas para designar as vendas.

Contrariamente ao modelo relacional, o modelo dimensional é assimétrico. Os projetistas de bancos de dados costumam chamar este modelo de Star Join Schema, por causa de seu diagrama que se assemelha a uma estrela com uma tabela dominante no centro, cercada por tabelas secundárias ou auxiliares que são exibidas em padrão radial e conectadas por uma única junção à tabela central. A tabela central é chamada de **tabela de fatos** enquanto as outras tabelas são chamadas de **tabelas de dimensão**.

5.11.2. Modelagem Dimensional

Em alguns aspectos, a modelagem dimensional mostra-se mais simples, mais

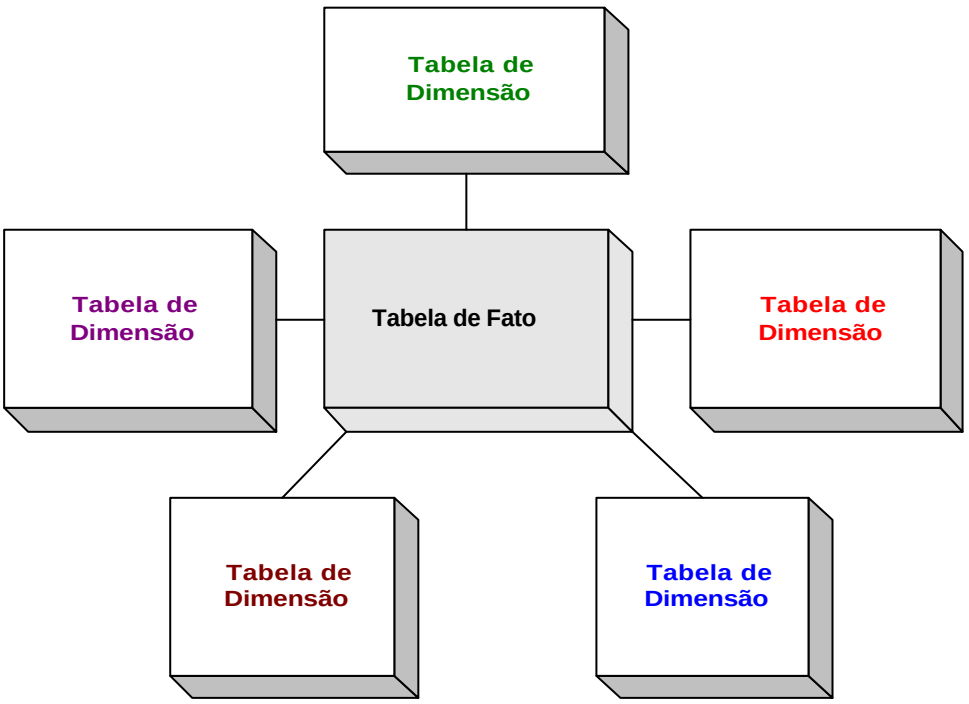

Figura 5.11.2.1. Modelo Dimensional contendo as tabelas de fato e de dimensão

expressiva e fácil de compreender do que a modelagem relacional. Mas sendo um conceito relativamente novo e por hora, não possuindo uma definição mais firme, é preciso que sejam apresentadas sua metodologia, técnica e algumas pistas sobre seu uso.

O conceito básico da modelagem dimensional apoia a sua técnica na visualização dos modelos de dados como conjuntos de medidas que são descritos pelos aspectos comuns dos negócios. Isto é um dos lados especialmente úteis nos processos de sumarização e re-arranjo, bem como na apresentação das visões dos dados que suportam as funções de análise. A modelagem dimensional, portanto, dá enfoque a dados numéricos, tais como valores, contas, pesos, balanços e ocorrências. Além disso, a MD tem como conceitos básicos, fatos, dimensões e medidas (ou variáveis), que serão descritos logo abaixo:

- **Fato** é a coleção de itens de dados relacionados, consistindo de dados de medidas e conceitos. Cada fato representa um artigo, item, transação, ou evento referente aos negócios da organização que podem ser passíveis de análise. No modelo dimensional, os fatos são implementados em tabelas nas quais todos os dados numéricos estarão armazenados. Tem-se como exemplo de fatos as vendas efetuadas e os estoques da empresa.
- **Dimensão** é uma coleção de membros ou unidades com o mesmo tipo de visão. No modelo dimensional, cada ponto referente a um dado de uma tabela de fatos está associado com um e apenas um membro de cada uma das múltiplas dimensões. Ou seja, são as dimensões que determinam o fundo contextual dos fatos apresentados, podendo ser mapeadas até entidades informativas e não-numéricas. Por esta razão, as dimensões funcionam como parâmetros no desempenho do processo analítico on line (OLAP). Além disso, uma dimensão pode agregar sob nomes distintos, um conjunto de itens com características e posições próprias, tais como meses e trimestres em relação a um período anual.

A estrutura do modelo dimensional pode ser melhor vista através de um cubo. Nele podem ser representadas, no mínimo, três dimensões. Na figura abaixo estão representadas três dimensões combinadas, **mercado**, **produto** e **tempo**, cujo objetivo é medir o volume de produção.

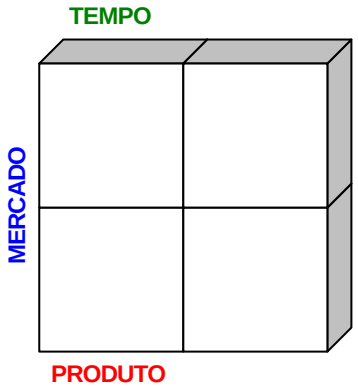

Figura 5.11.2.2. Cubo de Kimball mostrando três dimensões de um negócio

5.11.2.1. Operações do Modelo Dimensional

O modelo dimensional é desenhado, em princípio, para suportar a ferramenta OLAP e o processo decisório nos negócios. Para que seja incrementado o seu uso desta forma, são necessárias quatro tipos de operações, considerando-se a granularidade dos dados: Drill Down, Roll Up, Slice e Dice.

As operações drill down e roll up são usadas para fazer movimentos das visões para cima e para baixo ao longo da hierarquia das dimensões. Com o drill down o usuário é capaz de navegar entre altos níveis de detalhe dos dados, enquanto roll up pode habilitá-lo a aproximar-se ao nível mais sumarizado dos dados.

Operações como slice e dice promovem a navegação dos dados através da figura do cubo. Slice faz o corte do cubo, de modo que se possa dar enfoque sobre algumas perspectivas específicas. Dice é a operação que permite a rotação sob outras perspectivas, de modo que o usuário possa fazer uma análise mais acurada dos dados apresentados.

Há vários desafios na concepção do Data Warehouse é um dos maiores dentre eles é saber como os usuários processam suas consultas. Porque a rapidez no

desempenho das consultas é fundamental para a análise de dados e rapidez diz respeito, quase sempre, à estrutura e ao volume dos dados. Os requerimentos de consolidação dos dados, como pré-cálculo e pré-agregação, são indispensáveis à agilidade na performance das consultas, por isso, pré-calcular e armazenar dados antes que as mesmas sejam realizadas podem reduzir consideravelmente o número de registros a serem recuperados e manter consistente e rápido o desempenho. Neste caso, o uso de operações como o drill down, por exemplo, que possibilita ao usuário se movimentar ao longo dos níveis da hierarquia de uma dimensão, pode resultar na escolha de um caminho melhor na consolidação ou pré-cálculo dos dados.

5.12. Área de Preparação do Data Warehouse

Diferente dos processos de transações on-line, cujo propósito é a captura de grandes taxas de dados em mudanças e suas adições, o Data Warehouse tem por objetivo organizar grandes volumes de dados estáveis para facilitar a análise e a recuperação das informações. Portanto, algumas considerações e atividades sobre o projeto de criação do Data Warehouse vão mostrar-se tão necessárias quanto as operações de extração, limpeza e refinamento de dados na preparação e carregamento dos dados no DW.

Uma das atividades mais importantes de projeto de DW é a criação da área de preparação. Ela inclui a confecção das tabelas que vão conter as chaves primárias e secundárias e as tabelas que comportarão dados em transformação. Outros tipos de tabelas também podem ser requeridos visando conciliar dados extraídos de diversas fontes ou até mesmo referências cruzadas sobre informações, estas com a finalidade de ajudar a identificar entidades comuns tais como o registro de clientes em sistemas que utilizam chaves diversas. Uma grande variedade de tabelas temporárias pode ser necessária para intermediar a transformação dos dados. É preciso ressaltar que o projeto da área de preparação de dados dependerá de fatores como a diversidade das fontes de dados, o grau de transformação e carregamento requeridos na organização dos mesmos, além de sua própria consistência.

O projeto da área de preparação de dados também deve conter tabelas que possuam esquemas idênticos aos das tabelas alvos do Data Warehouse e dos processos

usados para extrair dados das fontes que os refinam e transformam, além, é claro, da seqüência de etapas que fazem o carregamento dos dados para o DW. Pode ser que estes processos sejam apresentados sob forma de pacotes DTS (Data Transformation Services) ou de documentos de manuais de instruções.

5.13. CONCEPÇÃO DE TABELAS

A próxima atividade mais importante na criação do Data Warehouse é a concepção das tabelas de fato e de dimensão e o estabelecimento de índices para campos-chaves em todas as tabelas. A definição da modelagem das mesmas tem sido a preocupação mais constante dos pesquisadores. Atualmente o modelo relacional vem cedendo espaço para o modelo dimensional que se mostra mais eficiente no armazenamento e consultas de dados.

Nas tabelas do modelo dimensional, juntos, o conjunto de dimensões e as medidas que lhes são associadas, compõem o que chamamos de fato. O processo de identificar e procurar agrupar dimensões e suas medidas sob requerimentos específicos dos usuários é primeiro passo para a obtenção de melhores respostas às consultas pelo DW.

As tabelas dimensionais armazenam descrições textuais que ajudam a definir os componentes das dimensões dos negócios. Uma tabela de dimensão pode possuir muitos atributos ou campos que se destinam a descrever os itens de uma dimensão. Assim, os atributos típicos de uma dimensão chamada produto incluem uma descrição sucinta (10 a 15 caracteres) e descrição longa (30 a 60 caracteres) do produto, como também sua marca, categoria, seu tipo de embalagem e tamanho.

Uma tabela de dimensão deve possuir uma chave primária, cuja especificação seja capaz de garantir sua integridade referencial. A tabela de fatos também possui uma chave primária que é chamada de chave composta ou chave concatenada, formada pela combinação de chaves externas. Todas as tabelas de fatos detém chaves compostas o que leva à conclusão que as tabelas de fatos possuem relacionamentos muitos-para-muitos, de outra maneira, são tabelas dimensionais.

5.14. Preparação de Dados

Dados para serem usados em um Data Warehouse devem ser extraídos de fontes. Refinados e formatados em seguida, para tornarem-se consistentes até serem transformados para entrarem no esquema do DW. O local onde isto ocorre chama-se Área de Preparação de Dados ou Staging Area. A área da preparação de dados é a base de dados relacional dentro da qual os dados ao serem extraídos das fontes, transformados com formatos comuns, checados por sua consistência e por sua integridade referencial estarão prontos para o carregamento nos Data Warehouses.

A área de preparação de dados e os bancos de dados do DW podem ser combinados em algumas implementações desde que as operações de refinamento e transformações não venham a interferir com o desempenho ou com as operações de serviço aos usuários finais do Data Warehouse.

Por causa da diversidade de dados proveniente de muitas fontes de carregamento on line para sistemas de transações, considerar apenas o desempenho nas operações em fontes de dados é raramente uma opção. O banco de dados relacional usado para preparação de dados deve, portanto, mostrar-se poderoso em termos de capacidade de manipulação e transformação dos mesmos.

Depois do carregamento inicial do Data Warehouse, a área de preparação de dados é usada em bases em andamento, a fim de preparar a atualização dos dados. Na maior parte dos sistemas de Data Warehouses estas funções de preparação são desempenhadas periodicamente, freqüentemente marcadas para minimizar o impacto na performance dos sistemas de fontes de dados operacionais.

O uso da área de preparação que é separada das fontes de dados e do próprio DW, promove o efetivo gerenciamento deste último. Visando transformar conjuntos de dados em sistemas de fontes pode interferir com a performance das transações do tipo OLTP. Ainda mais se os sistemas legados não possuírem capacidade efetiva de transformação. O fato é que o re-atamento da inconsistência entre dados extraídos de fontes diversas raramente pode se cumprido, até que os dados estejam coletados em

bases de dados comuns, sobre as quais os erros de integridade podem ser facilmente identificados e retificados.

A área de preparação deveria, em princípio ser isolada daquela dos dados do Data Warehouse para preservar a integridade deste e ainda permitir sua função primária de preparação da informação para apresentação e suporte de acesso a clientes. Isto porque muitas operações de Data Warehouses requerem consultas realmente sofisticadas e o processamento de grandes volumes de dados e seu refinamento pode interferir com estas operações.

Finalmente, a área de preparação funciona como uma espécie de banco de dados relacional que serve como uma área de trabalho geral para operações de preparação de dados. Assim, a mesma conterá tabelas que relacionarão chaves de fontes de dados à chaves do Data Warehouse, tabelas de transformação de dados e muitas tabelas temporárias. Também conterá processos e procedimentos, tais como os pacotes de serviços de transformação de dados (Data Transformation Services - DTS), que extraem dados a partir de fontes de dados dos sistemas.

5.14.1. Extração de Dados

A extração de dados é mais uma das operações mantidas pelo DW. Extrair consiste em retirar dados de sistemas operacionais, tanto no período de criação do Data Warehouse quanto nas fases de atualizações dos mesmos. Pode-se visualizar a extração como uma operação simples se a fonte de dados, no caso, vier a ser uma base de dados do tipo relacional, ou pode tornar-se mais complexa se os dados forem provenientes de sistemas operacionais heterogêneos. Porém, o objetivo deste processo é buscar de todas as fontes possíveis e confiáveis os dados, trazendo-os para um depósito comum a todos, enquanto são formatados com consistência e de modo a tornar seguro e direto o seu carregamento dentro do Data Warehouse.

As falhas mais freqüentes na atividade de extração de dados consistem em erros de validação e inconsistências. No primeiro caso a ocorrência se dá a partir da fonte dos dados, local onde devem ser processadas as correções pela implementação da detecção e checagem de erros do sistema operacional de onde os dados são extraídos. As

inconsistências, por sua vez, também são processadas quando os dados são extraídos de fontes diversas, mas utilizam diferentes sistemas de codificação para o mesmo tipo de dado. A solução plausível é utilizar tabelas de tradução a fim de conciliar essas diferenças seja durante o período de extração, seja durante as operações de transformação dos dados. A extração dos dados usa, na maioria das vezes, as ferramentas listadas a seguir:

- Transacções SQL
- Consultas Distribuídas
- DTS
- Aplicações de Linha de Comando
- Back-up
- Scripts ActiveX
- Comando *Bulk Insert* para carregamento a partir de arquivos de texto

Algumas implementações permitem que seja utilizada a replicação para se copiar dados dos sistemas de fontes para a área de preparação.

5.14.2. Refinamento e Transformação

Juntas, as atividades de refinamento e transformação de dados procuram combinar dados provenientes de fontes diversas durante o processo de extração. Através de operações como a formatação e a incorporação das chaves, pode-se obter a conciliação técnica dos dados que vai permitir o incremento da performance do Data Warehouse. Estas atividades acontecem na área de preparação dos dados e são completadas antes que se faça o carregamento dos mesmos.

A atividade de refinamento dos dados é baseada em três componentes: validação dos dados, limpeza e manipulação de erros.

A validação dos dados consiste em um determinado número de checagens, incluindo a legitimação de valores para um atributo (domain check), ou seja, verificar se o atributo é válido no contexto de uma linha ou de linhas relacionadas de uma, ou entre várias tabelas e a checagem do relacionamento entre as linhas de uma tabela e entre outras tabelas válidas (foreign key check).

A limpeza dos dados é o processo que os torna mais significativos para o DW. O exemplo mais comum de limpeza de dados a sincronismo das informações sobre o nome e o endereço de um cliente que, em geral, armazenadas em múltiplas locações, tendem a perder esta característica.

Por manipulação de erros é entendido o processo que indica o que fazer com dados abaixo dos padrões de perfeição instituídos para o DW. Assim, da perspectiva do modelo, os dados podem ser rejeitados, armazenados para reparo posterior em área específica, ou passado para o Data Warehouse apesar de suas imperfeições. Neste último caso, os metadados devem incluir direções sobre a qualidade esperada dos dados (tipos de erros) e sobre a freqüência dos erros contidos nos dados. Durante a atividade de refinamento dos dados pode-se executar também os procedimentos de verificação de consistência.

A operação de transformação é um passo crítico nos esforços de desenvolvimento do DW. Há duas grandes decisões a serem tomadas neste ponto: como capturar os dados e fazer a formatação e a incorporação das chaves. Na verdade, nesta operação, os dados devem ser direcionados para um objetivo exclusivo, o que significa

dizer a geração de um plano documentando os passos a serem seguidos para se capturar dados e direcioná-los da fonte ao alvo principal, ou seja, é simplesmente adicionar mais metadados.

5.14.3. Carregamento de Dados

Depois de refinados e transformados em uma estrutura consistente com os requerimentos de um Data Warehouse, os dados estarão prontos para serem carregados. Carregar dados significa popular as tabelas segundo o esquema do Data Warehouse e verificar sua preparação para o uso. Alguns métodos mais comuns são:

- Transações SQL
- DTS
- bcp utility

A atividade de carregamento não só envolve a apresentação dos dados aos usuários finais, como também a transferência de vastos volumes de dados a partir de fontes de sistemas operacionais, a preparação do banco de dados que servirá como área de preparação de dados e o arranjo das áreas das tabelas do banco do próprio Data Warehouse. Estas últimas operações são sempre desenvolvidas em períodos de baixa utilização do sistema.

5.14.4. Verificação de Integridade dos Dados

Depois de carregados os dados, entram as ações de verificação da integridade referencial entre as tabelas de dimensão e as tabelas de fato. Como regra geral, deve-se entender que a verificação de integridade serve para garantir a consistência dos dados no resultado de uma consulta. A verificação dos dados assegura ao usuário que ao navegar em uma estrutura dimensional, ele terá garantida a mesma resposta para qualquer pesquisa que realizar em qualquer tabela. Para que isto ocorra é preciso assegurar que todo registro é relacionado apropriadamente em uma e outra tabela indiferentemente, ou seja, que para cada registro de uma tabela de dimensão corresponderá um registro em uma tabela de fato. Mas como toda regra tem exceção, a decisão de reforçar a verificação de integridade pode levar à rejeição de um dado ou à sua inclusão se for

decidido que é melhor manter um dado inconsistente do que se privar dele. Neste caso, o relaxamento na verificação da integridade é perfeitamente compreensível, desde que sejam conhecidas as suas conseqüências pelos usuários do DW.

As atividades acima descritas normalmente se desenvolvem na fase de criação do Data Warehouse, entretanto, algumas como o carregamento podem ocorrer também durante a manutenção do mesmo.

5.14.5. Atualização

A correção de dados é obviamente necessária tanto que ocorram modificações em hierarquias, status, propriedades, etc. Dados incorretos acarretam erros ao sistema do Data Warehouse como um todo e contrariamente ao que se crê, nem todas as mudanças que ocorrem nos data marts que originam informações para o DW, são repassadas a tempo de atualizar as informações.

O gerenciamento das atualizações deve ser previsto no projeto, de modo que não se permita a perda ou até mesmo pior, o uso de dados que podem induzir o usuário a um falso resultado na análise.

5.14.6. Outras Atividades

Algumas atividades podem ser descritas como parte do projeto do Data Warehouse, mesmo que não sejam aquelas que se destinam à preparação dos dados propriamente dita. Estas atividades não são desenvolvidas na fase de implementação do projeto, mas são complementares ao mesmo, uma vez que sem elas, o sentido do Data Warehouse pode ser perdido.

- **Consulta -** Consulta é um termo abrangente que incorpora todas as atividades de pesquisa de dados a partir dos data marts, incluindo consulta ad hoc, elaboração de relatórios, suporte de aplicações à decisões complexas, busca por modelos e mineração de dados. Consultas nunca ocorrem na área de preparação, mas nos servidores de apresentação dos dados do DW.

- **Realimentação ou Feedback** - A atividade de re-alimentação dos dados se dá logo que o Data Warehouse é posto em uso. Uma vez que os dados são capturados e refinados na área de preparação e armazenados em sistemas legados, é desejável que se mantenha sua qualidade e consistência em verificações e trocas constantes de seus valores para os usuários. A re-alimentação do sistema é a atividade que se encarrega de manter o "suprimento" de dados no nível certo e com as características adequadas.

- **Auditoria** - De tempos em tempos é importante saber de onde vêm os dados e como são desenvolvidas as outras atividades que os tornam valiosos como peças de informação. As técnicas de auditoria foram criadas para se passarem nas fases de extração e transformação dos dados que ocorrem na área de preparação. A partir de lá, os registros são criados e posteriormente ligados diretamente aos dados reais auditoriados, para que o usuário possa usá-los como informações adjacentes a qualquer época.

- **Segurança** - Todo Data Warehouse possui o mesmo dilema: como publicar seus dados para tantos quantos forem os usuários que necessitem deles, com as interfaces mais fáceis de se utilizar, e mesmo assim, protegê-los das invasões dos hackers e espiões industriais? O desenvolvimento da Internet tem amplificado drasticamente este dilema, mostrando que os construtores de DW têm uma preocupação que vai além de prover seus clientes com dados de boa qualidade e consistência. Esta atividade ganha peso em fases preparatórias do projeto de Data Warehouse, uma vez que encarna a principal fobia do staff da organização, a invasão do sistema.

- **Back Up e Recuperação de Dados** - Desde que o Data Warehouse é o fluxo constante de dados que partem dos sistemas legados para os data marts e, eventualmente, para os desktops dos usuários, a questão que se impõe como real é quando fazer os instantâneos que assegurarão a recuperação dos dados em caso de falência do sistema. Adicionalmente, pode ser ainda mais complicado fazer o back up e a recuperação do metadado que azeita, por assim dizer, as engrenagens das operações do Data Warehouse. No entanto, são ameaças reais como a perda de dados por falhas nos sistema que obrigam os analistas a levarem em conta mais esta preocupação sob pena de um desastre maior.

5.15. Serviços de Apresentação

O propósito do Data Warehouse é expor informações sobre negócios em uso no processo de tomada de decisão. Pode o DW, portanto, conter milhares de dados que, sem o devido refinamento, tornam-se sem utilidade como ferramenta de análise. Estas ferramentas podem variar de simples relatórios a sofisticados algoritmos de mineração de dados.

Alguns exemplos:

- Relatórios Pré-definidos
- Processos Analíticos On-line
- Mineração de Dados ou *Data* Mining
- Programas de Interfaces de Aplicação

5.16. Serviços de Análise ao Usuário Final

Refere-se às ferramentas com interface gráfica que permitem ao usuário final executar uma série de atividades relacionadas à recuperação e otimização de consultas a uma base de dados.Podem ser utilizados em consultas a registros em forma de texto ou grid; pela visualização da representação gráfica das etapas de execução da própria consulta; na análise da performance de índices; como auxílio on-line das consultas; e na execução de scripts ou stored procedures.

5.17. Tipos de Data Warehouses

Considerando-se a função analítica dos dados, Tanler (1988) descreve cinco opções de projetos de Data Warehouse, ressaltando que as tabelas de dados que cada um possui, em princípio, devem ser adequadas às necessidades e ao perfil da organização as quais são propostas:

Star Schema (Estrela) – Projeto originado na indústria de vendas por atacado, cujos negócios são analisados através de dimensões simples (produtos e mercados), sendo ainda estáticas no que se refere a números e estrutura, em relação ao tempo. Oferece diversas vantagens, entre os quais, a utilização de uma única tabela de fatos e uma só tabela de dados por categoria, o que garante a reutilização de metadados. Além disso, sua performance é melhorada em decorrência da emissão de uma única declaração em SQL para a tabela de dados a cada consulta. A tabela correspondente ao modelo é conforme abaixo:

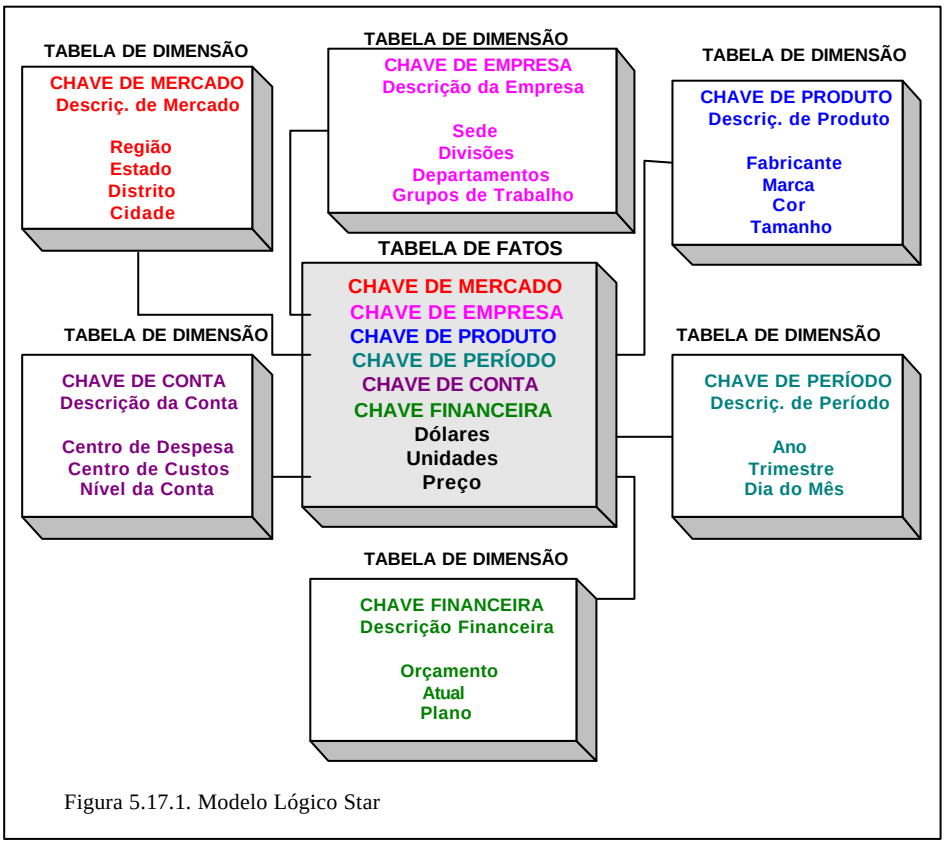

Figura 5.17.1. Modelo Lógico Star

Parcial Star (Estrela Parcial) – É uma espécie de expansão do modelo anterior, só que neste caso, é permitido às organizações (setor bancário, farmacêutico, de vendas por catálogo e de seguros) manter uma grande quantidade de entidades por nível de agregação. Neste modelo existem diversas tabelas que podem ser combinadas e particionadas por diversos níveis de agregação, em cada dimensão. As chaves geradas são utilizadas em cada dimensão e por causa disso, e do tamanho reduzido de seus índices, tem alta performance. Este tipo de tabela é mostrado a seguir:

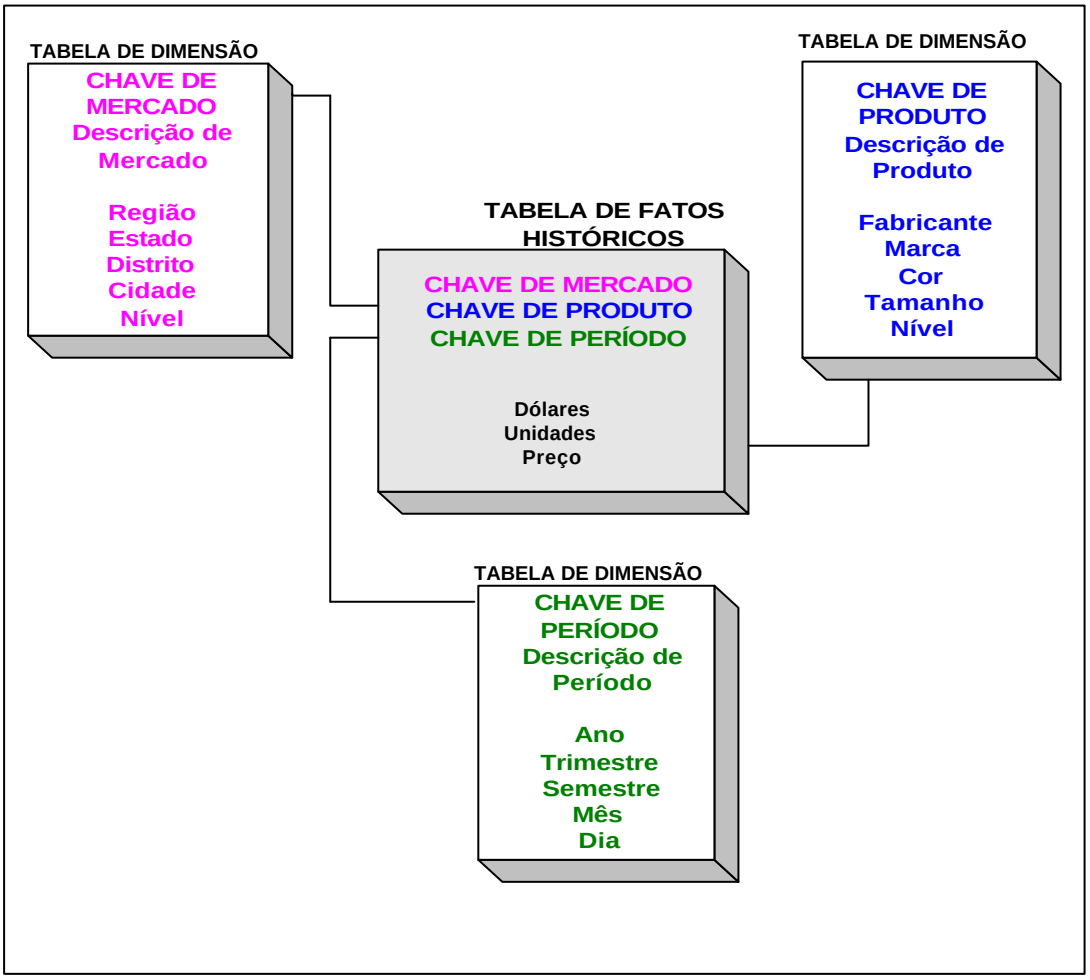

Figura 5.17.2. Modelo Lógico Estrela Parcial

Dimension Partitioning (Particionamento de Dimensão) – Combina princípios do modelo estrela parcial com o modelo estrela, do qual está mais para ser uma variação. Possui uma tabela de fatos para cada categoria, mas sua lógica é mesclada a várias tabelas de dimensão que, por sua vez, são particionadas a nível de sumarização. Sua desvantagem é justamente o armazenamento de informações de detalhes e de sumarização em uma tabela, o que pode reduzir sua performance.

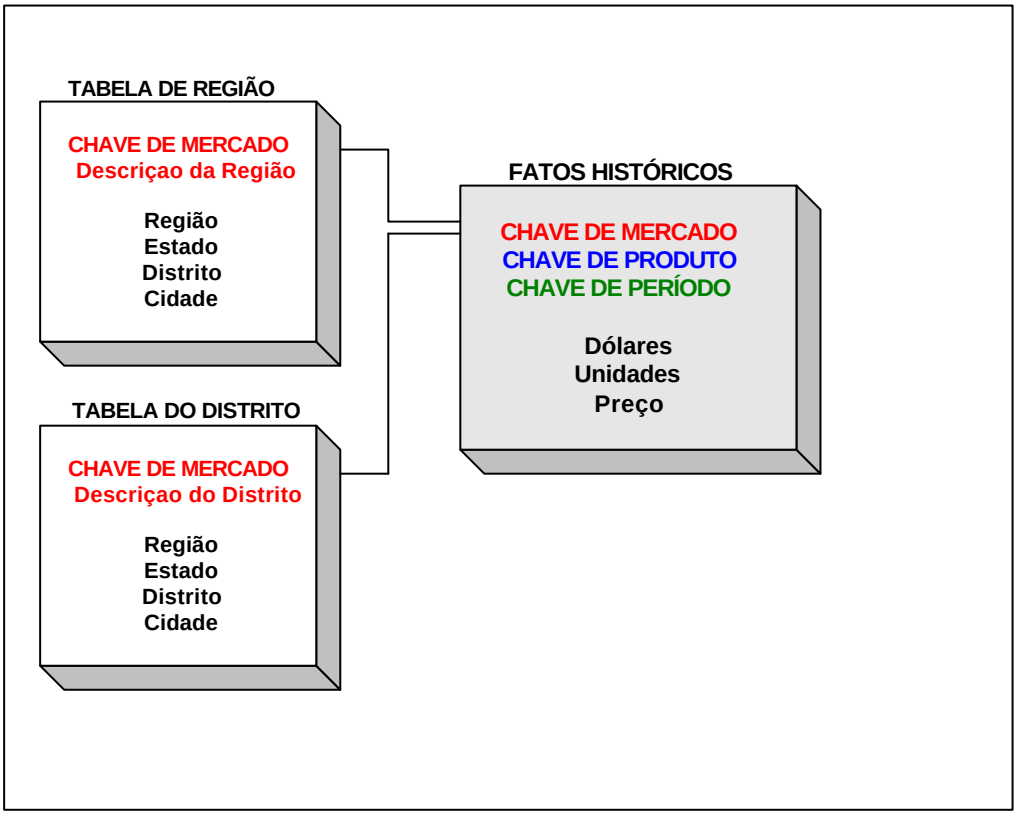

Figura 5.17.3 Modelo Lógico de Particionamento de Dimensão

Fact Partitioning (Particionamento de Fatos) – É uma variação do modelo estrela parcial, mas também aplica alguns princípios do modelo Star Schema. Possui uma única tabela por dimensão que é mesclado com diversas outras tabelas de fatos (que existem no interior de cada categoria) e são particionadas por níveis de sumarização, o que significa dizer que o mesmo fato pode pertencer a diversas tabelas, as quais vão pertencer a uma tabela de maior dimensão. A desvantagem deste modelo é a redundância de descrição dos fatos.

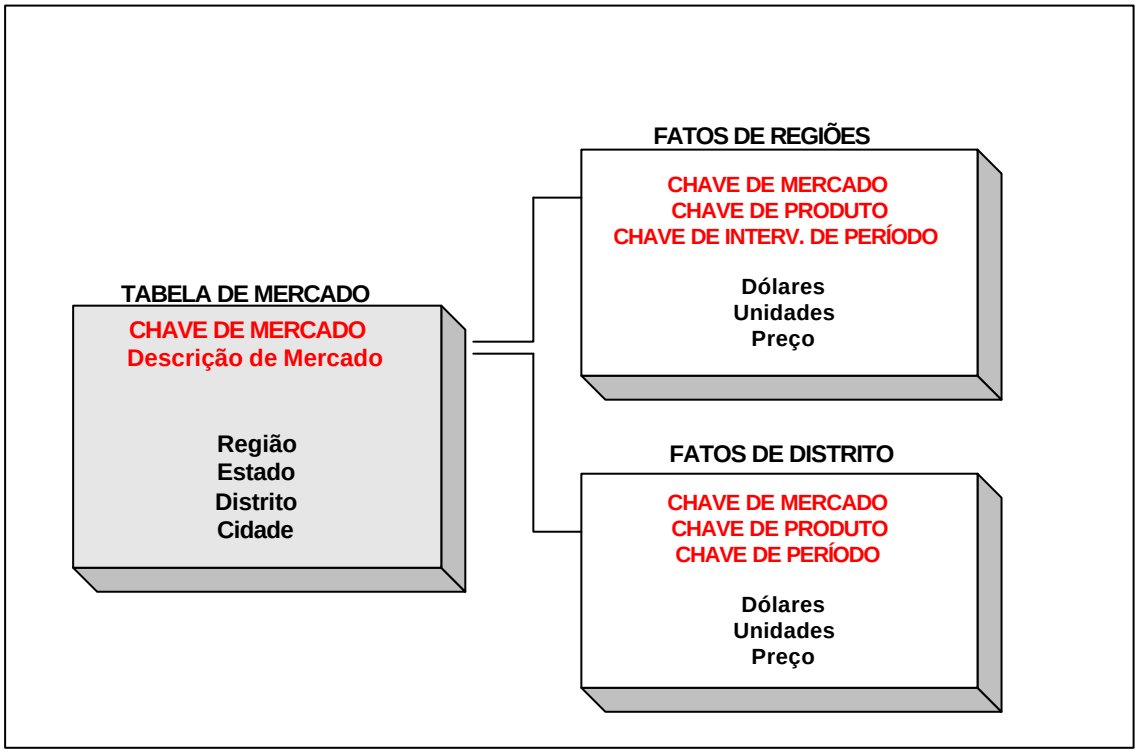

Figura 5.17.4. Modelo Lógico de Particionamento de Fatos

Snowflake Model (Modelo Floco de Neve) – São desenvolvidos de modo a suportar a capacidade de armazenamento da descrição de produtos, mercados e empresas em um único local através do uso de uma combinação de normalização de bancos de dados, visando manter a integridade dos dados e reduzindo seu número para alcançar uma performance mais elevada. Estes modelos incorporam grandes tabelas de dimensão que possuam uma junção lógica com tabelas de fato. Suas tabelas principais de dimensão são parecidas com as tabelas de dimensão do modelo estrela, entretanto, suas colunas de atributos contém chaves para as tabelas outrigger em vez de descrições de texto.

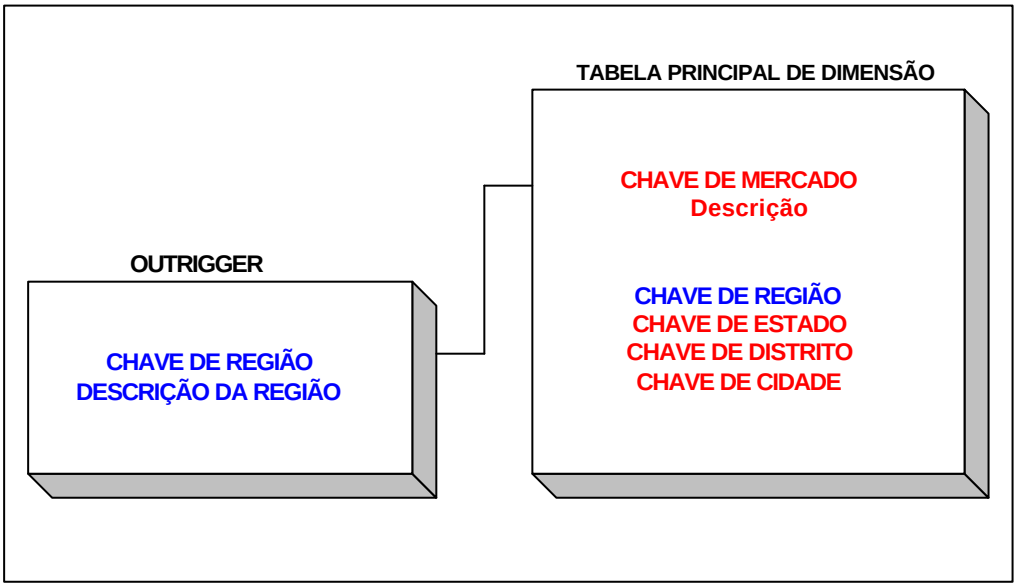

Figura 5.17.5. Modelo Lógico Snowflake

5.18. A Escolha dos Modelos

São dois os esquemas básicos mais usados na modelagem dimensional: Snowflake e Star.

O esquema Star é o termo comum também usado para emprestar conotação ao modelo dimensional, tanto que os projetistas de Data Warehouse têm usado este esquema como sendo o significado do próprio modelo dimensional. A principal razão é que o esquema Star, como seu próprio nome sugere, possui como estrutura resultante a aparência de uma estrela, da qual o modelo lógico é o esquema físico. Assim sendo, este esquema tem uma tabela central, de grande tamanho, chamada de tabela de fatos e um conjunto de tabelas menores, chamadas de tabelas de dimensão, distribuídas em padrão radial, em volta da tabela maior.

O modelo dimensional é tipicamente identificado por fatos e dimensões, logo que se toma conhecimento dos requerimentos dos negócios da organização. Inicialmente, como visto antes, a modelagem dimensional é melhor o representada pela aparência de uma estrela em cujo centro figura a tabela de fatos rodeada por um nível de diversas tabelas. Snowflake é o esquema que resulta da decomposição deste primeiro nível de tabelas em uma ou mais dimensões, as quais, algumas vezes podem vir a ter níveis hierárquicos formado por relacionamentos do tipo muitos-para-um (M:U). É derivado do esquema Star e de fato, como seu próprio nome sugere, tem o aspecto de flocos de neve, mas com uma estrutura de fácil visualização de níveis hierárquicos e a vantagem de economizar espaços nas tabelas por armazenar dados repetitivos em sub-dimensões, por assim dizer. Entretanto, ressalta-se que este não é o melhor critério para a decisão na escolha de esquema para o modelo dimensional, pois a estrutura do modelo snowflake pode ser tornar complexa e desconfortável para os usuários na hora da consulta. Além disso, outro critério de peso é o uso da ferramenta OLAP que prioriza as respostas rápidas em consultas ad hoc.

Se por um lado os dois esquemas apresentados parecem, à primeira vista, muito diferentes um do outro, não é novidade que também podem apresentar similaridades. Na modelagem dimensional, por exemplo, eles podem apresentar a mesma notação para definir entidades, relacionamentos, atributos e chaves (primária e estrangeiras). A

conclusão a que se pode chegar é que ambos possuem, como tudo mais, seus pontos fortes e suas fraquezas e que cada um pode ser usado apropriadamente em situações diversas.

No capítulo seguinte serão descritas a garimpagem e a utilização on-line de armazenamento e análise de dados em Data Warehouse, tendo em uso as ferramenta Data Mining e OLAP (On-line Analytical Processing).

DATA MINING E FERRAMENTA OLAP

No capítulo anterior foram apresentados os esquemas mais comuns de armazenamento de dados para modelos de Data Warehouses dimensionais. A maior parte dos esquemas pode ser utilizada sem problemas em projetos de DW e depende inteiramente dos analistas e desenvolvedores a escolha do esquema mais adequado ao projeto.

Neste capítulo serão vistos dois aspectos de extrema importância para o armazenamento de dados para DW: a mineração e a disponibilidade on-line dos dados. Serão mostradas também sobre quais aplicações de negócios devem estar alinhados estes dois fatores, assim como suas aplicações.

6.1. Estratégias de Aplicação de Dados

Uma vez compreendido que os processos da empresa devem ser automatizados em um sistema que releve a importância das informações sob forma de aplicações coletivas, é preciso assimilar o fato de os processos que compõem o sistema são inumeráveis, tanto em termos de monitoramento quanto aos aspectos de compartilhamento de informações.

Há dois tipos básicos de aplicações sobre os quais se apoiam as estratégias do sistema organizacional:

As Aplicações para Negócios – Referem-se àquelas aplicações que fazem os negócios da organização serem operacionalizados ao longo dos dias, semanas, meses, etc., os quais, literalmente, apoiam a empresa.

As Aplicações Sobre os Negócios – São as aplicações que analisam os negócios, seja interpretando o que ocorre, seja decidindo ações a serem tomadas para o futuro. Estas não estão relacionadas com as operações diárias que fazem funcionar a empresa, mas são críticas como fatores de competitividade para as mesmas.

Estas últimas constituem a espinha dorsal do Data Warehouse e é sobre elas que são constituídas as estratégias de um projeto de Data Warehouse.

6.2. Estratégias de Informação sobre Negócios

Existem três conjuntos individuais e interligados de estratégias dentro de um projeto de Data Warehouse. Eles são estabelecidos após a definição clara dos fatores de sucesso de uma organização e da documentação dos objetivos que determinarão o alcance de tais fatores. Os conjuntos de estratégias mais importantes são:

- **Estratégias de Informações** – Reúne-se em torno do levantamento da necessidade de informação e da análise de fatores críticos de sucesso da organização. Esta estratégia está ligada ao armazenamento e utilização de informação dentro de bancos de dados. Ela define parâmetros da escolha de dados, como também os programas aplicativos a serem utilizados nas funções de análise e geração de relatórios.

- **Estratégias de Bancos de Dados** – Envolve os tópicos que irão garantir a qualidade das informações contidas em bases de dados, assim como a escolha dos software e hardware para o sistema de gerenciamento destas bases. A escalabilidade é um de seus pontos mais importantes, pois está relacionada a quantidade de usuários, bem como a complexidade da análise dos dados.

- **Estratégias de Acionamento** – Relativa à capacidade técnica e compartilhamento dos dados entre usuários. Leva em conta, ainda, a facilidade de acesso aos dados disponibilizados e o modo de recebimentos destes dados pelos usuário móveis e remotos.

Para Perkins (1996), estratégia em Data Warehouse significa juntar dados baseados nos requerimentos fundamentais dos negócios, prevendo, principalmente, metodologia e ferramentas baseadas em computadores, com flexibilidade e capacidade para desenvolvimento.

6.3. Estruturação das Informações

Para Boar (1999), a estratégia da informação tem a ver com a construção de vantagens. Nisto estão incluídas a composição e a manutenção dos fatores fundamentais e dominantes que vão sustentar os negócios.

As idéias básicas que vão suportar a estratégia de informação do Data Warehouse são manuseamento, precisão e aquisição adequada de informação. Alguns fatores são de extrema importância para a estratégia de informação do DW:

- Maximização da qualidade, do acesso e do compartilhamento de dados
- Eliminação da redundância não-planejada de dados
- Simplificação da interação de inter-aplicações
- Segurança na padronização dos dados
- Maximização do ciclo de vida produtiva dos dados
- Desenvolvimento de novas aplicações pela aceleração através da reutilização das fontes de dados
- Criação de centros de excelência em gerenciamento de dados, com a finalidade de proteção dos dados engajados

Na composição da estratégia de informação do Data Warehouse as estruturas básicas mais relevantes são:

- **Estrutura Física** – Bases físicas de dados, nas quais todos os dados do Data Warehouse são armazenados, junto com metadados e processos lógicos, com fins de refinamento, organização, enquadramento e processamento dos dados detalhados.

- **Estrutura Lógica** – Contém metadados, incluindo as regras e processos lógicos da empresa. Também passam pelas atividades de refinamento, organização, enquadramento e processamento, mas não podem conter dados atualizados. Pelo contrário, contém apenas a informação necessária para que se possa acessar os dados onde que estejam armazenados.

- **Data Mart** – Subconjunto sobre todos os dados do Data Warehouse. Tradicionalmente, suportam os elementos organizacionais tais como departamentos, regiões, funções, etc. Os data marts são partes interativas da empresa e fazem a ligação entre as diversas partes lógicas do Data Warehouse, alimentando-o como se fossem simples Data Warehouse, eles próprios.

6.4. Estratégias de Bancos de Dados

Dados históricos, apesar de estáticos, continuam a ser o melhor objeto de análise de uma situação. Para as organizações em processo de tomada de decisão (o que pode ocorrer diariamente), o conhecimento das informações passadas reforça a confiança nas decisões a serem tomadas no presente e para o futuro.

Tome-se como exemplo, a reflexão histórica em vários períodos de tempo sobre investimentos financeiros de uma empresa. Embora os dados não sofram alterações, eles fornecem o retrato fiel de como a organização se comportou face a mudanças sofridas no contexto mercadológico, sua postura em períodos de crise ou de crescimento econômico e em função do cenário no qual está inserida.

Em épocas mais antigas, eram os relatórios impressos, produzidos pelas gerências é que sustentavam os serviços de informações das empresas em avaliações ou processos decisórios. Com a melhoria na relação custo/benefício no uso da informática para o armazenamento de dados, o Data Warehouse, que sustentado por bancos de dados

relacionais, tecnologia do tipo on-line, tais como OLAP e técnicas de mineração de dados, tornou-se praticável e eficiente sob vários aspectos.

Em princípio, Data Warehouses foram construídos baseados no gerenciamento de bancos de dados fundamentados em modelos lógicos que são baseado em objetos, usados na descrição de dados em níveis conceitual e visual, tal como o modelo Entidade/Relacionamento.

6.4.1. Modelo Entidade/Relacionamento

Modelo baseado na percepção do mundo real, a partir da qual é constituído um conjunto de objetos ou *entidades* e seus *relacionamentos.*

É um modelo poderoso e muito usado em projetos de sistemas de processamento de transações em ambientes relacionais, do tipo OLTP. Caracteriza-se por ser uma técnica de modelagem que normaliza e automatiza as estruturas físicas dos dados. Mas, embora tenha contribuído de modo intenso na aquisição de grandes volumes de dados para bancos relacionais, este modelo não possui a habilidade requerida para consulta a dados.

Tendo sido desenvolvido para facilitar o projeto de banco de dados que permita especificar um esquema empresa, o modelo Entidade-Relacionamento ou E-R pode definir percepções com as quais o conteúdo do banco de dados deve estar de acordo. A percepção mais comuns é a normalização, que reduz a redundância dos dados e permite separar os itens de dados em tabelas distintas com junções por chaves, além, de simplificar as operações de atualização e inserção, pois quando ocorrem modificações, elas ocorrem apenas em um registro das tabelas a que se relacionam.

Entretanto, a maioria dos casos de uso deste modelo pelas empresas, mostrou a proliferação quase ilimitada do número de tabelas para representar os processos de negócios. Tome-se como exemplo, um modelo E-R para todas as atividades referentes apenas às vendas de produtos. Um diagrama para representá-la não possui menos do que algumas centenas de tabelas, conectadas a outras centenas, sem mencionar suas junções. O resultado visual final supera, em termos de volume, as perspectivas de análise do usuário final. Ou seja, diagramas do tipo E-R são úteis na otimização de desempenho de

transações on-line, pois são projetados para serem vistos em pequenas seções, tal como acontece nos sistemas OLTP, mas reduzem em muito a visão analítica de tabelas necessária aos processos decisórios.

6.4.2. Modelo Relacional

Segundo Korth (1996), o Modelo Relacional aplicado a bancos de dados, possui perspectiva histórica relativamente nova em relação a outros modelos de armazenamento de dados.

E. F. Codd, então pesquisador da IBM, o idealizou visando fazer frente às deficiências de complexidade, flexibilidade e dependência de métodos de armazenamento físico de dados dos bancos e sistemas de gerenciamento desenvolvidos na década de 60.

O modelo relacional possui um projeto técnico e lógico, cujo objetivo principal é remover a redundância dos dados e baseia-se em duas ramificações da matemática, a Teoria dos Conjuntos e o Predicado Lógico.

Seu nome foi tirado a partir da correspondência entre o conceito de tabela e o conceito matemático de relação. Nele estabelece-se uma coleção de tabelas, cada qual contendo suas próprias linhas e designadas por um nome único. A terminologia utilizada por Codd serve para referir-se aos relacionamentos entre os conjuntos de valores relacionados nestas linhas, além de explicar as relações entre uma série de tabelas, que não são ordenadas, mas que podem ser manipuladas usando-se operações não-procedurais. Sua estrutura assemelha-se a do modelo Entidade/Relacionamento.

Atualmente o modelo Relacional é utilizado em funções de armazenamento, atualização e/ou recuperação de itens de dados compartilhados de forma simples ou múltipla, em sistemas operacionais e de transações em bancos de dados. Mesmo assim, boa parte dos produtos e ferramentas associados ao modelo Relacional tem apresentado significativa limitação quanto à eficiência nos processos mais complexos, como os decisórios, o que tem levado ao desenvolvimento e introdução maciça do modelo dimensional nas atividades de gerenciamento de informação em processos decisórios.

6.4.3. Modelo Dimensional

Com a crescente necessidade de expandir os bancos de dados para análise, a modelagem dimensional tem se mostrado mais adequada aos sistemas de Data Warehouse. Em primeiro lugar, porque este tipo de modelo é o que apresenta mais habilidade para refletir a consistência e a lógica dos negócios. Em segundo lugar, porque ele se mostra capaz de relevar o processo de tomada de decisão dada sua forma inusitada de apresentação de dados referentes aos negócios.

Kimball (1996) descreve o Modelo Dimensional ou Star Schema como sendo o modelo cujos principais componentes são as tabelas de fato e as tabelas de dimensões. Seu diagrama assemelha-se a uma estrela, com uma grande tabela no centro rodeada por tabelas menores, exibidas em padrão radial.

Neste modelo, a tabela central ou de *Tabela de Fatos* é a maior de todas as tabelas existentes. É onde se localiza o "grão" ou foco da consulta sobre a área de medição do negócio. Esta tabela é simples e altamente desnormalizada. Por definição, ela armazena fatos reais sob forma de atributos numéricos, aditivos, sumarizados e com relacionamento do tipo muitos-para-muitos. Sua chave primária tem apenas uma coluna chave por dimensão, mas contém um conjunto de duas ou mais chaves estrangeiras que representam sua junção ou interseção com as chamadas tabelas de dimensão.

O Data Warehouse de uma organização pode possuir um bom número de tabelas de fatos que, separadas, representam diferentes processos dos negócios da empresa, tais como *Pedidos, Inventário, Orçamento,* etc. Estas tabelas de fatos vão possuir ligações com tantas tabelas de dimensão quanto possível.

Kimball et al. (1998), descrevem as *Tabelas de Dimensão* como conjuntos de tabelas de acompanhamento que lhes provêem informações textuais, ou descritivas e que servem de base para a integridade referencial para qualquer tabela de fato com a qual possuem junção. Segundo os mesmos autores, a maioria das tabelas de dimensão contém atributos (campos) textuais, que agrupam e limitam as buscas dentro do Data Warehouse.

Ainda, na modelagem dimensional o banco de dados é visto como um cubo, cujas dimensões fatiadas (uma, duas, três, ou até mais), possuem perspectivas muito

simples e navegação eficiente. Cada uma das arestas do cubo representa uma área de atividade da empresa. As coordenadas definidas e combinadas pelas arestas mostram as medições do negócio da organização. Além disso, sua abordagem "cubista" simples pode ser implementada em bancos de dados relacionais, melhorando sua visualização.

Para melhor exemplificar a modelagem dimensional, tomemos como exemplo uma organização que oferece *bens* em vários *mercados* e mede seu desempenho ao longo do *tempo*. O negócio como um todo pode ser imaginado como um cubo cujos pontos interiores representam os local em que as medições do negócio armazenam as combinações de Bens, Mercados e Tempo.

O argumento final para a utilização técnica da modelagem dimensional associada ao modelo relacional de banco de dados é que ela põe em foco, de forma mais concreta e tangível, os dados, incrementando o desempenho nas atualizações através da normalização que, via de regra, é usada pelo modelo relacional.

6.5. Garimpagem de Dados para Análise

Nos últimos tempos o acúmulo e o armazenamento de dados em formato eletrônico sofreram um aumento considerável dentro de organizações dos mais variados tipos e tamanhos. Para que se pudesse gerenciar mais facilmente as tecnologias de informação, impondo-lhe mais significado e aumentando as vantagens competitivas nos negócios, foi necessário formalizar o discernimento e a capitalização do valor nas consultas a bases de dados. Para isto, surgiram três alternativas plausíveis. Na primeira aumentou-se o poder de armazenamento dos multiprocessadores. Na Segunda, houve um decréscimo nos custos destes processadores. E por último, surgiu a preocupação com uma orientação mais adequada quanto ao refinamento dos dados utilizados em processos decisórios.

Ao processo evolucionário originado para extrair ou garimpar dados, deu-se o nome de data mining.

Data mining, é também conhecido como Knowledge Discovery in Databases - KDD e empregado em campos tais como:

- **Medicina** – Para conhecimento dos efeitos de medicamentos, análise de custos hospitalares, análise de seqüência genética, predições, etc.
- **Finanças** – Previsões e rotação de estoques, avaliação de crédito, detecção de fraudes, etc.
- **Marketing/Vendas** – Análise de produtos, previsões de vendas, análise de comportamento de consumidores, etc.
- **Aquisição de conhecimento em Pesquisas Científicas** – Condução e orientação de pesquisa e experimentos.
- **Engenharia** – Diagnóstico de sistemas automotivos, detecção de falhas, etc.

Segundo Frawley et al. (1996), data mining é a extração da informação não-trivial, previamente desconhecida e potencialmente utilizável de dado, que inclui um número de diferentes tipos de abordagens técnicas, tais como clustering, sumarização de dados, regras de classificação, dependência de redes de trabalho, análise de mudanças e detecção de anomalias.

Para Holshemier e Siebes (1994), data mining é a procura para relações e padrões globais que existem em bancos de dados grandes, mas que são encobertas entre a vasta quantidade de dados e seus diagnósticos. Estas relações representam um valioso conhecimento sobre o banco de dados e seus objetos, ainda, determinam se o banco de dados é um espelho fiel do mundo real o qual procura refletir.

Sob outro prisma, a analogia do processo da extração e garimpagem de dados pode ser descrita como o uso de uma variedade de técnicas para identificar os grãos mais valiosos de informação ou de conhecimento em dados, extraindo-os, de tal modo, que podem ser postos em uso como apoio às funções de decisão, predição, prevenção e estimativa. Dados são freqüentemente volumosos e como aparentemente o uso através de sua extração direta se prova de baixo valor, conclui-se que é a informação escondida nos dados que é útil.

O desenvolvimento do data mining foi impulsionado ao ponto de permitir a seus usuários, a navegação sobre dados, em tempo real. Hoje, a maturidade de suas técnicas associada ao excelente desempenho das ferramentas de bancos de dados relacionais e dimensionais, impuseram a prática desta tecnologia ao ambiente de Data Warehouse, servindo de apoio em tanto que ferramenta de refinamento de dados para negócios

6.5.1. Processos de Garimpagem de Dados

Data mining faz uso de ferramentas de software e hardware para descobrir modelos e situações padronizadas sobre conjuntos de dados. O processo em si, segue a metodologia utilizando uma amostra constituída de um conjunto de dados extraídos de uma fonte qualquer e que vai servir para desenvolver a representação de uma estrutura ótima de dados para a aquisição de conhecimentos. Uma vez que a estrutura é construída, estende-se a mesma técnica aplicada a todos os outros conjuntos de dados coletados (de quaisquer que sejam seus tamanhos), pois intui-se que tais conjuntos tenham estrutura similar à da amostra utilizada. Esta analogia é aplicada objetivando-se extrair o valor máximo de informações que uma coleta que os grandes volumes de dados possam apresentar.

O Data mining é apoiado por três importantes linhagens: a *Estatística Clássica*, que serve de base para o grande número de tecnologias de extração de dados; a *Inteligência Artificial* ou IA, que é constituída a partir de fundamentos heurísticos; e a chamada *Machine Learning*, melhor descrita como sendo a associação entre as duas primeiras, onde se aliam os conceitos fundamentais da estatística com técnicas heurísticas da inteligência artificial.

As técnicas de data mining são resultados de longos processos de pesquisa sobre produtos orientados ao armazenamento de grandes volumes de dados sobre negócios. São passíveis de atualização e podem ser implementadas em sistemas novos, sem maiores problemas. Outra vantagem concernentes à estas técnicas é que ao serem implementadas em processadores paralelos, as ferramentas de data mining são capazes de analisar volumes maciços de dados em minutos, o que significa um incremento de rapidez nas funções de predição de dados.

As abordagens técnicas mais comuns em data mining são:

- **Associações**: Referem-se aos relacionamentos mais significativos entre itens e dados armazenados. Seu objetivo detectar tendências no grande número de transações que possam ser usadas para entender e explorar os padrões de comportamento dos dados. Ex: Na área da saúde pode-se definir correlações entre um ou mais sintomas e determinados tipos de doenças.

- **Padrões Seqüenciais**: Tal como ocorre no item anterior, pode-se identificar eventos que ocorram de tempos em tempos, tais como as doenças epidemiológicas, como as que ocorrem durante as estações mais frias do ano.

- **Séries Temporais Similares**: As séries similares armazenadas na base de dados e que variam de forma semelhante ao longo de períodos de tempo são identificadas. Ex: Os preços de produtos provenientes de diversos fornecedores, que variam de maneira quase idêntica.

- **Classificação e Regressão**: Aqui, utilizam-se os dados armazenados para criar modelos de comportamento de variáveis. Um grupo inicial de registros tomado como padrão, é denominado de "conjunto de treinamento". Os demais registros são classificados a partir destes padrões. Com um padrão de comportamento das variáveis definido, pode-se determinar quais registros estão fora deste padrão e conhecer, assim, o próprio distanciamento deste padrão, confirmando e explicando, de certa forma, a verificação de algumas anomalias encontradas posteriormente;

- **Clusterização**: A informação disponível é segmentada em conjuntos definidos e homogêneos, baseando-se em atributos específicos. Este conceito já é conhecido em diversas áreas, porém em data mining foi especializado a fim de permitir a sua aplicação em itens não numéricos. Neste tipo de algoritmo não é informado ao sistema os tipos de classes existentes, ficando a cargo do computador descobrir classes a partir das alternativas encontradas na base de dados;

- **Detecção de Desvios**: É possível encontrar anomalias na base de dados através da computação das informações, quando se compara-as com regras e padrões definidos, que não podem ser quebrados, tais como a realização de determinados procedimentos que somente podem ser feitos em indivíduos do sexo feminino, exames preventivos de câncer de mama.

6.6. Data Warehouse e Ferramenta OLAP

Nos últimos anos a tecnologia de Data Warehouse cresceu rapidamente a partir de um conjunto de idéias correlacionadas dentro de uma arquitetura para entrega de dados aos usuários finais das organizações e a tecnologia OLAP foi a abordagem dimensional de suporte à decisão que mais contribuiu para sua evolução.

Um Data Warehouse é tradicionalmente mantido separado dos bancos de dados operacionais da organização por duas razões importantes: A primeira é que esta ferramenta tem por composição uma expressiva e variada quantidade de dados, tornando-a ponto de convergência e integração de múltiplas fontes.

A outra razão refere-se ao fato de que um Data Warehouse deve ser capaz de suportar o chamado processo analítico *on-line* ou OLAP (On-Line Analytical Processing), cujos requerimentos funcionais e de performance diferem daqueles do processo de transações on-line (OLTP - On-Line Transaction Processing), normalmente suportados pelos bancos de dados operacionais.

Portanto, é principalmente nestes dois aspectos, que o Data Warehouse vai diferenciar sua estrutura daquelas de outros tipos de sistemas, fazendo com que, em vias normais, ele deva ser implementado em uma máquina separada.

6.6.1. OLAP - Conceito

O Processo Analítico On-line ou OLAP visa suprir a necessidade de análise de dados on-line. Ele é definido como uma abordagem dimensional de suporte à decisões, pois pode responder em tempo hábil à grande quantidade de questões que surgem no decorrer desta atividade.

O conceito de função de análise de dados on-line foi introduzido por E.F. Codd, em estudo publicado na década de 80, onde ficaram estabelecidos critérios originais para diferenciar as ferramentas analíticas deste sofisticado processo, diferenciando-o totalmente dos processos mais simples de pesquisa e relatório.

A filosofia de base de OLAP está altamente alinhada com o conjunto de 12 critérios proposto no artigo de E. F. Codd, intitulado *"Providing OLAP (On-Line Analytical Processing) to User-Analysts: na IT Mandate"*, publicado em 1993, para a Arbor Software, agora chamada Hyperion Solutions Corporation, a fim de representar formalmente um conjunto de padrões de comparação para sistemas de suporte à decisões.

Codd, que foi pesquisador na área de bancos de dados durante duas décadas (60 a 80), também é creditado como sendo o idealizador do modelo de banco de dados relacional sobre o qual as regras são, em princípio, aplicáveis, embora também sejam amplamente utilizadas como parâmetros para compor modelos multi-dimensionais de Data Warehouse.

As regras criadas por Codd como diretrizes para OLAP provocaram controvérsia em razão do suporte de sua pesquisa ter sido feito por vendedores de sistemas deste tipo. Por causa disso, muitos pesquisadores as consideram mais como sendo parte de uma brochura do que propriamente um artigo científico. Alguns anos após a publicação do artigo onde as designava como regras, Codd as redefiniu e reestruturou em quatro grupos que veio a chamar de Features.

Pendse (1997), define OLAP como a classe de software com tecnologia para capacitar o Data Warehouse a operar como ferramenta de suporte no apoio a decisões.

A ferramenta OLAP pode auxiliar analistas, gerentes, executivos ou qualquer outro tipo de usuário final a acessar de modo rápido, consistente e interativo as informações que dão dimensão ao negócios da organização.

6.6.2. Funções de OLAP

Além das pesquisas de Codd (1993) e Pendse (1990), estudos posteriores sobre a ferramenta OLAP, cujo objetivo era segmentar os sistemas projetados de pesquisa e análise de dados visando torná-los mais eficientes e melhor distribuídos, estabeleceu valores mais informativos e específicos através da existência de quatro diferentes elementos:

- **Garimpagem de dados** – Atividade que se refere ao tipo mais complexo da ferramenta OLAP. Nele são introduzidos sofisticados algoritmos (algoritmos com ramificações de decisões, redes neurais, lógica oculta e algoritmos genéricos), visando identificar as relações entre itens de dados.

- **Análises Multidimensionais** – É a atividade que permite aos usuários acessar o Data Warehouses em uma única dimensão, mas com flexibilidade suficiente a poderem direcionar a análise iniciada para outras dimensões, onde a capacidade analítica torna-se mais expandida, através da navegação em camadas.

- **Análise Estatística** – É citada como fórmula matemática, por apresentar a capacidade de redução de uma grande quantidade de dados a uma simples relação. Suas técnicas estatísticas permitem a geração de modelos na projeção de comportamento de vários elementos (clientes, índices de vendas, aceitação de produtos no

mercado, etc.) que interagem com uma organização e que servem como base para tendências e relações históricas.

- **Consulta e Geração de Relatório** – Atividade que permitem aos usuários formular consultas a bancos de dados, sem que tenham que interagir com qualquer linguagem de programação, por mais simples que estas se apresentem.

6.6.3. Operações de OLAP

A funcionalidade do processo OLAP é, na maior parte dos casos, caracterizada pela análise multi-dimensional e se apoia no conceito de navegação através de dados. Além disso, ela é suportada por atividades como:

- Cálculos e modelagem dimensionais de dados, aplicados através de hierarquia e/ou a todos os membros
- Tendência de análise sobre períodos de tempo seqüenciais
- Repartição de subconjuntos para visualização em tela
- Vazamento para níveis mais profundos de consolidação
- Busca através de níveis mais detalhado de dados
- Rotação de novas comparações dimensionais nas áreas visualizadas

A ferramenta OLAP é implementada em máquinas com arquitetura cliente/servidor para multi-usuários. Este modelo de arquitetura é capaz de oferecer consistência na rapidez das respostas em qualquer tipo de busca, a despeito do tamanho ou da complexidade do banco de dados.

Por outro lado, ela também ajuda a sintetizar de forma personalizada as informações referentes aos negócios da organização para fins de comparação, assim como pode tornar-se apto a apresentar uma análise histórica dos dados em cenários projetados para simular situações com maior conjunto possível de diversidade.

6.6.4. Cliente OLAP

São aplicações para usuários finais que requerem elementos provenientes de servidores OLAP e são também capazes de prover exibições bi ou multi-dimensionais, seleção, alterações feitas pelos usuários, posicionamentos de diversos tipos, cálculos, etc., para visualização e navegação sobre dados. Os clientes do tipo OLAP podem ser simples como uma folha de um documento inteiro, ou sofisticadas como as apresentações de modelos de análises, de vendas, de marketing, etc.

6.6.5. Servidor OLAP

Os servidores OLAP tem por finalidade a entrega de aplicações do tipo warehouse, ou seja, aplicações que requeiram dados históricos, projetados e derivados. Um servidor OLAP é uma ferramenta de alta capacidade, utilizada por múltiplos usuários e projetada especificamente para dar suporte ao acesso a estrutura de dados que operem em plataformas multi-dimensionais.

Servidores OLAP diferem de servidores do tipo OLTP em muitos aspectos, pois eles tem por missão conduzir o gerenciamento crítico de dados que são acessados dentro de um processo interativo e analítico de investigação, enquanto os servidores OLTP provêem acesso a dados em consultas muito mais simples. Os robustos mecanismos de cálculo dos servidores OLAP podem combinar o acesso padrão com as poderosas ferramentas de análise de dados.

A estrutura multi-dimensional é preparada de modo que cada item de dado alocado e acessado seja baseado na interseção das dimensões que são definidas por seus membros. O projeto do servidor e a estrutura dos dados são otimizados para que facilitem a rapidez das consultas ad hoc, de maneira que a recuperação das informações receba orientações de flexibilidade e transformação de dados brutos sejam firmados em relações sobre fórmulas.

Um servidor OLAP também é capaz de compor a plataforma física que processa informações multidimensionais, de modo a poder distribuir respostas rápidas e consistentes aos usuários finais, ou até mesmo, fazer a ocorrência das estruturas de seus

dados em tempo real, partindo de bancos de dados relacionais ou de quaisquer outros tipos, e até mesmo oferecendo escolha entre ambos.

6.7. Aptidões da Ferramenta OLAP

Segundo pesquisas da Hyperion Solution Corporation (2000), servidores OLAP precisam estar aptos a:

- Reduzir grandes volumes de dados em escala e atender numerosos usuários concorrentes
- Possuir tempo de resposta consistente e rápido para permitir análises interativas a velocidade de pensamento
- Integrar metadados próprios com aqueles dos bancos de dados relacionais do Data Warehouse
- Migrar de dados calculados e resumidos para dados detalhados armazenados em bancos de dados relacionais
- Possuir mecanismos de cálculos para funções matemáticas robustas em computação de dados derivados tais que: agregação, cálculo de matrix, cálculo de dimensão cruzada, fórmulas OLAP-ware e cálculos procedurais)
- Integrar, sem emendas, dados históricos, calculados e derivados
- Fornecer ambiente multi-usuário (read/write) para dar suporte a atividades de análise do tipo *what-if* e a requerimentos de modelagem e planejamento
- Estar habilitado a ser organizado e usado em tempo hábil, a facilitar sua adoção e manter um custo eficiente
- Gerenciar usuários e prover segurança a acessos robustos de dados
- Oferecer vasta gama de ferramentas de visualização e análise para suporte de diferentes comunidades de usuários.
- matrix, cálculo de dimensão cruzada, fórmulas OLAP-ware e cálculos procedurais)

6.8. OLAP e Data Mining

Diferentemente das funções OLAP, explicadas anteriormente, e utilizadas no Data Warehouse pelos usuários finais em sofisticadas e específicas questões de processo decisório, o data mining utiliza-se de algoritmos de padrões de reconhecimento e de aprendizagem para identificar relações entre itens de dados e oferecer resultados sobre tendências e agrupamentos de grandes quantidades de informações que lhes são entregues.

Juntas, as duas tecnologias se integram ao Data Warehouse como elementos de prospecção, refinamento e análise de dados, destilando e dinamizando a visualização de informações que irão capacitar a organização a aplicar ações mais dinâmicas a seus planos estratégicos e táticos.

6.9. Ferramentas de Prospecção e de Análise

Ferramentas de prospecção e de análise on-line de dados são elementos fundamentais para sistemas que provêem informações com qualidade, consistência e precisão aos usuários de Data Warehouses. Juntas, as ferramentas de Data mining e OLAP têm o objetivo comum de manter a lapidação de dados brutos e sua oferta em tanto que informação de valor aos tomadores de decisão da empresa.

METODOLOGIA PARA DATA WAREHOUSE

Organizações comerciais estão, cada vez mais, ordenando e colocando em uso grandes volumes de dados sobre seus negócios. Dados, estes, obtidos através de transações operacionais e sistemas legados e utilizando ferramentas do tipo Data Warehouse.

O trabalho feito sobre a tecnologia do Data Warehouse dá ênfase ao armazenamento e ao acesso de dados extraídos de várias fontes, sejam internas, como por exemplo os sistemas operacionais, sejam externas que podem ser inúmeras e variadas.

A estrutura fixada para o projeto de DW não possui outra alternativa senão somar dados coletados, refiná-los, armazená-los e oferecer acesso com navegação simples, a fim otimizar as consultas feitas para análise em processos decisórios.

Em teoria, as informações oferecidas pelo Data Warehouse são completas e confiáveis, pois de modo geral são a alta gerência e os analistas de negócios que utilizam as suas aplicações através de redes de trabalho do tipo Cliente/Servidor ou Intranets.

Em outros tipos de aplicações, os Data Warehouses trabalham com sistemas operacionais, em uma espécie de reiteração, para desempenhar funções sobre informações direcionadas, especializadas, mas sem caráter analítico, tais como re-ordenamento de inventário, por exemplo.

Originariamente, o Data Warehouse foi idealizado como uma espécie de container onde se depositariam “informações sobre informações”. Atualmente, ele é uma ferramenta tecnológica potente, projetada e construída com orientação às matérias específicas dos negócios. E as empresas, hoje, organizam e fazem uso de múltiplos Data Warehouses, com diferença táctil em dados em muitos níveis de sumarização para uso, principalmente, em se tratando de suas comunidades de usuários finais.

A produção do sistema do Data Warehouse consiste em muitos componentes, que incluem análise de clientes e ferramenta de relatório, extração de dados de sistemas legados e transações em sub-sistemas, além de gerenciamento de metadados. Entretanto, o componente mais ativo e importante do DW é o servidor do Sistema de

Gerenciamento do Banco de Dados Relacional (SGBDR), usado para armazenar vastas quantidades de informação e apresentá-las com rapidez e confiabilidade.

Durante os últimos dez anos, uma percentagem significativa de dados corporativos vem migrando para os bancos de dados relacionais, que por sua vez, vêm sendo usados em áreas de operações e controle, com ênfase particular em ambiente especializado de processos de transações on-line (OLTP), especialmente sobre manufatura e comercialização de produtos. A capacidade tecnológica dos processos OLTP mostram-se capazes de satisfazer as demandas de aplicações requeridas pelo Data Warehouse em função do estado-da-arte em que se calcula, ocorram mais de mil transações por segundo. Isto é o que, em suma, obrigaria o Data Warehouse a priorizar requisitos como o desempenho e a concorrência entre usuários, em detrimento outras coisas consideradas menos importantes.

Entretanto, mesmo sobre todos aspectos positivos que um Data Warehouse possa apresentar para a organização, é sobre a função de análise de dados que recaem as maiores expectativas das empresas.

Tradicionalmente, a organização fatual dos dados é estabelecida em torno dos conceitos de negócios, tais como produtos ou serviços, clientes, pedidos, aplicações financeiras, etc. Por esta razão, faz-se necessária a distinção das capacidades de um Data Warehouse entre as de ferramentas OLTP e OLAP que normalmente o compõem. Considerando-se as prioridades de cada uma das ferramentas em função das atividades que desempenham, é compreensível que a escolha de OLAP no projeto de DW seja a mais apropriada. Isto não quer dizer que também não possam ser utilizadas as habilidades da ferramenta OLTP, apenas em estágios onde não hajam requisitos de análise de dados.

Outro ponto a ponderar é o uso da tecnologia relacional de bancos de dados pela maioria dos projetos de Data Warehouse. Nos últimos anos têm havido investimentos maciços no emprego de bancos de dados dimensionais, por demonstrarem estes grande poder de facilitar a visualização dos dados agregados da empresa e rápido acesso à informações estratégicas, tão necessárias à tomada de decisões, já que os bancos de dados relacionais, sem que se possa tirar-lhes o devido mérito, são agregações de dados

cuja maior qualidade é o acelerado processo de transações operacionais, que não necessariamente servem como base para requisitos analíticos. Por isso, vem crescendo o número de adeptos da modelagem dimensional de dados.

Ainda, em relação ao projeto, é preciso deixar claro que o Data Warehouse por maior que seja a sua importância, não pode ser visualizado como um todo em relação aos dados que abriga. Para se falar de DW é preciso mencionar também a importante colaboração dos data marts que normalmente o compõem. Sem os data marts, o DW não seria mais do que o acúmulo desordenado de dados que dificilmente se prestariam ao seu maior papel: o de fornecer informações com qualidade e consistência.

Neste capítulo serão vistos, primeiramente, o emprego do DW como apoio à função de análise, componentes e proposta de modelagem.

7.1. Uso, Processos e Componentes de DW

Com freqüência, Data Warehouses contém dados históricos que são coletados de uma variedade de fontes tais como processos de transações on-line, sistemas legados, arquivos de textos e spreadsheet. Combinar estes dados e refiná-los, procurando acuidade e consistência enquanto se organiza-os para facilitar a eficiência nas consultas requer múltiplos componentes e operações como aqueles vistos anteriormente, além de uma metodologia básica utilizando-se a modelagem dimensional que, por sua vez, será vista no presente capítulo.

Na maior parte das organizações, os relatórios são uma forma distinta de auxiliar a análise da situação da empresa. Através deles pode-se obter exatidão e rapidez ao se examinar os aspectos fundamentais dos resultados nas consultas.

Como coleção de tecnologia evolutiva de decisão e suporte, o Data Warehouse envolve, em sua maior parte, atividades dentro de processos e sub-processos básicos que são relacionados ao armazenamento e acesso a dados:

- **Extração** – É conhecido como sendo o primeiro passo para se introduzir dados no ambiente do Data Warehouse. Extrair significa, ao mesmo tempo, a leitura e a compreensão da fonte da qual o dado

é proveniente, além, é claro, da atividade de cópia de partes do dado necessárias à sua apresentação para trabalhos futuros.

- **Transformação** – Fase do processo que inclui os seguintes passos:

Refinamento: Limpeza dos dados pela correção ortográfica dos mesmos, resolução de conflito de domínios (tais como nomes de cidades com códigos postais incompatíveis), arranjos a serem feitos sobre a falta de elementos de dados e análise e associação em grupos de unidade com caracteres sintáticos para padronização de formatos.

Purificação : Faz a seleção de campos a partir de dados de sistemas legados, mas que não são úteis ao Data Warehouse.

Combinação : Estabelecimento de parcerias exatas sobre valores chaves ou sobre desempenho indistinto de acordos sobre atributos sem chaves, os quais incluem a busca por equivalentes textuais de códigos de sistemas legados.

- **Carregamento** : Após as atividades de extração, refinamento e transformação, os dados devem ser carregados dentro do Data Warehouse. Outros pré-processos tais como checagem de integridade, ordenação apropriada de itens, sumarização, agregação, construção de índices, tabelas e caminhos de acesso podem, entre outros, podem ser requeridos na fase de carregamento de dados.

- **Indexação** : Uso de um número para selecionar um elemento a partir de uma lista, vetor, coleção ou outro tipo de seqüência.

- **Checagem** : Inspeção dos dados é feita sobre o conjunto de dados que foram carregados no sistema. É o tipo de processo que deve estar presente sobre todas as categorias de dados, de modo que satisfaça os padrões de qualidade dos mesmos. Todos os valores

> devem ser relatados e consistentes com as séries similares de valores que os precedem. As publicações dos dados que passaram pelo processo de checagem podem ser comunicada aos usuários finais, desde que a qualidade e as possíveis atualizações ocorridas sejam asseguradas antes de sua publicação como fato consolidado.

Resumidamente, estes processos são associados a três componentes básicos do Data Warehouse que podem facilitar o acesso aos dados, tornando-os mais rápidos, consistentes e flexíveis. São eles:

- **Diretório de Informação** – Consiste no gerenciamento de informação sobre o qual os usuários buscam informações técnicas e de negócios sobre a organização. Este diretório funciona como uma espécie de assistente que auxilia os usuários a entenderem que existem fontes utilizáveis de informações com as quais eles podem entender o significado dos negócios, criar e fazer consultas, além de analisar e relatar situações específicas.

- **Entrega de Dados** – É o componente usado para distribuir os dados para os data marts, através de servidores descentralizados do Data Warehouse. Seu conteúdo, também chamado de coleção de dados, assim como seu calendário de distribuição, são definidos e usados pelo assistente do Diretório de Informações.

- **Acesso a Dados** – Consiste de uma base de dados do tipo *middleware*, somada às ferramentas de busca e de análise dentro do Data Warehouse, em geral, produzidas pela ferramenta OLAP.

7.2. Proposta de Modelagem

A idéia central desta pesquisa é propor a modelagem dimensional de dados, seguindo o esquema Star e a abordagem bottom up. Embora sejam muitas as formas de aplicações dos esquemas, abordagens e componentes para Data Warehosue, a modelagem aqui sugerida possui qualidades gerais que permanecem indiscutíveis nas aplicações em quase todas as áreas, sejam elas numéricas, financeiras, de marketing, etc.

A modelagem dimensional, como vista anteriormente é aquela que melhor disponibiliza dados em caráter analítico. Através dela, pode-se ter as mais variadas e concretas visões que se possa esperar deste tipo de dados. As dimensões servem de apoio à navegação de dados, enquanto seus usuários percorrem as informações de que necessitam.

O esquema Star é a opção natural, por mostrar-se menos complexo na construção, apresentação e compreensão das dimensões. Através dele, as visões são facilitadas e em número exato, de modo a se obter domínio dos dados armazenados.

A abordagem bottom up mostra-se condizente com uma modelagem da qual se espera melhor gerenciamento de dados, por possuir baixa complexidade e melhor relação custo/benefício, pois em geral, seus investimentos em hardware são baixos. Além disso, resulta em design mais simples do que aquele do Data Warehouse global.

Na modelagem proposta foram considerados três caracteres fundamentais a serem vistos:

- **Visualização multidimensional dos dados** - As visões multidimensionais dos dados são inerentes aos modelos que representam os negócios e raramente estão limitadas a menos de três dimensões por modelo. Isto proporciona ao cenário corporativo, a interação entre áreas e suas atividades, pois as organizações raramente são olhadas sem uma percepção segmentada em cubo (slice and dice), principalmente se existirem processos analíticos com requerimentos de informações agregadas e cruzadas. Agregações básicas são desempenhadas sobre certos tipos de dimensões (produto, cliente e canais de distribuição), enquanto dimensões com cálculos mais

complexos são reunidas em outras dimensões. Em geral, as dimensões de medidas computam proporções e médias. A discrepância ou variantes são incluídas ao longo da dimensão do cenário da própria organização. O modelo complexo baseado em desempenho histórico é usado para montar as previsões do cenário. Finalmente, firmeza na consistência do tempo de resposta para qualquer tipo de consulta é a chave para estabelecer servidores com habilidade para fornecer visualizações multidimensionais de dados.

- **Capacidade de compreensão do modelo** - Compreender o modelo significa dizer que ele pode ser visto e interpretado em qualquer uma de suas dimensões. Além disso, pode servir para formar a idéia do todo dos dados armazenados e possuir facilidade de transferência de uma dimensão a outra.

- **Inteligência da abordagem proposta -** Tempo, qualidade e consistência são fatores que, reconhecidamente, vão provar a eficiência de qualquer aplicação analítica. Tempo é uma dimensão única, por ser seqüencial em sua própria natureza e ser capaz de entender e manejar dados em função deste fator é imperativo nas situações emergenciais. Aliás, a própria natureza dos negócios é, por si mesma, avaliada em função do tempo. Mas hierarquia do tempo não é para ser usada do mesmo modo que as outras hierarquias no Data Warehouse. Conceitos sobre comparações de períodos iguais (ano a ano, dia a dia, etc.) devem ser facilmente definidos pela no balanceamento do tempo. Pode-se regulamentar o modo como o fator tempo vai ser computado para aspectos como inventários, médias de unidades vendidas, previsões de vendas, em dias, semanas, meses, trimestres, semestres e anos, além de seus fatores de atualização. exemplo, este mês relaciona-se aos meses seguintes ou anteriores, assim como acontece com o

período relativos a anos. A qualidade diz respeito à propriedade do em ser aceito pelo valor que carrega em si. Qualidade também se define em função da necessidade que se tem de determinado dado, ou seja, uma informação pode ser crucial em processos de decisão, mas se não tiver o tratamento correto quando extraída de uma fonte operacional ou externa, compromete-se todo o seu significado para a análise. Consistência é o estado de firmeza ou concordância que um dado apresenta em todas as dimensões onde possa ser armazenado. Através da consistência, sabe-se que uma informação é ou não verdadeiramente utilizável em processos decisórios.

7.3. Resultados Esperados

Pretende-se, com a proposta de modelagem de um Data Warehouse dimensional, utilizar os conhecimentos teóricos abordados nos capítulos anteriores, enquanto se propõe um projeto, em linhas simples, dos mesmos e considerando um prévio conhecimento de bancos de dados, dos processos, preparação e atividades relativos à modelagem de um DW. Os resultados esperados são a construção de tabelas de fato e dimensionais para uma loja de departamentos, onde pode-se copiar o modelo para outras áreas como a de finanças, marketing, estoque, etc. comprovando assim, a efetividade de um Data Warehouse, bem como sua flexibilidade, consistência e desempenho em relação às funções de análise dentro da organização.

7.4. Etapas da Modelagem do Data Warehouse

As etapas do trabalho de avaliação da ferramenta OLAP são:

1. Selecionar uma ferramenta entre os modelos relacional e dimensional.
2. Estabelecer pontos de decisão, onde:
    a) Identifica-se o processo a modelar

   b) Determina-se a escolha do grão que vai popular as tabelas
   c) Compor as tabelas do modelo dimensional
3. Fixar os fatos mensuráveis guardados pelas tabelas.
4. Gerar chaves primárias e externas.
5. Criar índices para melhorar as consultas especializadas.
6. Composição de visões para as tabelas.
7. Criação de Diagramas para as tabelas de dimensão e de fato.

O próximo capítulo tem o objetivo de mostrar a construção simplificada e baseada na teoria apresentada nos demais capítulos. É preciso levar em consideração que não houve escolha de uma ferramenta específica de bancos de dados, pois entre as diversas que estão à disposição dos usuários no mercado, existem algumas de muito bom calibre tais como SQL Serve e Oracle. Entretanto, o objetivo do trabalho não é mostrar o desempenho de nenhuma delas, mas sim, colocar em prática o conjunto de orientações de vários autores para a construção de um Data Warehouse.

APLICAÇÃO DA METODOLOGIA

Existem duas abordagens para a criação de um sistema de Data Warehouse em uma empresa. No primeiro caso, o DW é projetado, desenvolvido e implementado para que sejam criados, em seguida, seus data marts. De outro modo, os data marts são implementados antecipadamente, pois servem de base para o Data Warehouse quando as informações contidas neles são reunidas. Qualquer que seja a abordagem, o projeto deve ser centralizado para que toda informação que diga respeito às atividades de negócios da empresa torne-se consistente e utilizável. A experiência mostra, no entanto que, somente os data marts que possuem projetos adequados ao próprio projeto do Data Warehouse são capazes de produzir as melhores respostas à consultas, como também relatórios mais sólidos, mesmo que suas informações residam em lugares diferentes.

Existe uma infinidade de pontos a ponderar quando se trata da criação de um Data Warehouse. Neste capítulo serão vistos de maneira simplificada as idéias centrais de um projeto de Data Warehouse a partir de alguns passo identificados como básicos para sua criação.

Neste capítulo são apresentados alguns aspectos relevantes para a modelagem de um Data Warehouse. Como mencionado anteriormente, supõe-se certo conhecimento em bancos de dados, sem o qual torna-se impossível a compreensão do presente trabalho.

A seguir serão mostradas também, as etapas e passos mais simples seguindo-se as orientações aqui mostradas.

8.1. Seleção da Ferramenta de Modelagem

Modelar para Data Warehouse é significativamente diferente de fazer o mesmo para sistemas operacionais. Em DW, qualidade e consistência são mais importantes do que a recuperação de dados ou tempo de resposta à consultas. Portanto, sendo o seu objetivo principal oferecer dados para análise, a preocupação com o acesso pelos usuários possui requisitos como o conhecimento das fontes, a transformação, agregação e controle do fluxo de dados, o que não acontece normalmente com sistemas operacionais.

As modelagens relacional e dimensional foram apresentadas neste trabalho como ferramentas disponíveis na construção do DW, entretanto, fica claro que seus diferentes propósitos em relação aos dados as tornam importantes para funções muito diversas. Por definição, o modelo relacional dá suporte a sistemas operacionais cujas atividades não estão relacionadas ao Data Warehouse. Por outro lado, quando já existente, este modelo serve de base na expansão de um sistema que incorpore as pretensões de analisar informações, pois qualquer ferramenta respeitável de modelagem de dados deve ter a habilidade suficiente para permitir a conversão dos mesmos, de uma notação para outra sem que se perder seu significado.

A modelagem dimensional, por sua estrutura que abriga tabelas de fato e de dimensões, além de permitir a utilização dos esquemas Star e Snowflake, possui mais aptidão para lidar com a análise de dados. Por exemplo, uma determinada cor ou símbolo pode ser usado para diferenciar uma tabela de fato de uma de dimensão, o que não acontece com as tabelas do modelo relacional que, ao se utilizar das notações de entidades, relacionamentos e atributos, além da cardinalidade dentro dos relacionamentos, torna senão difícil, impossível a distinção entre tabelas, a visualização imediata de seus relacionamentos e, portanto, seu uso para a análise de dados. Considerando que um modelo dimensional é representado visualmente como uma tabela de fato cercada por tabelas de dimensão, freqüentemente é chamado um esquema estrela (star schema). Assim, a modelagem dimensional e o esquema Star colocam-se como a melhor escolha em termos de ferramenta de apoio ao projeto de Data Warehouse proposto neste trabalho.

8.2. Pontos de Decisão

Algumas etapas importantes devem ser cumpridas em qualquer projeto de Data Warehouse. Aqui vão ser abordadas algumas delas de forma simples e objetiva.

A primeira etapa, define o ramo do negócio a ser modelado e os dados que lhe estão relacionados. A modelagem do negócio permite escolher entre segmentos comerciais tradicionais de produtos, incluindo fabricação, atacado e varejo, e a prestação

de serviços como os setores financeiro, de transportes, ou órgãos governamentais, por exemplo.

O passo seguinte dispõe sobre qual grão da tabela de fatos pode vir a ser relacionado ao negócio. Grão é o nível fundamental e atômico que representa cada processo na tabela de fatos e que determina a dimensionalidade do banco de dados a ser utilizado. Em um modelo dimensional, o grão tabela de fato normalmente é uma medida quantitativa do resultado do processo empresarial analisado. Como exemplo de grãos típicos têm-se as transações individuais e instantâneos diários, semanais ou mensais, onde muitos efeitos importantes das atividades da organização podem ser identificados.

Durante a terceira etapa, escolhe-se as dimensões que serão aplicadas como registros de tabelas. As dimensões mais utilizadas são: *Tempo, Produto, Cliente, Promoção, Almoxarifado, Tipos de Transações e Status*. Para cada dimensão escolhida são descritos diferentes atributos (campos), que preencherão cada tabela dimensional.

Por fim, escolhe-se os chamados fatos mensuráveis que irão popular cada registro de tabela de fatos. São considerados fatos mensuráveis típicos as quantidades numéricas aditivas como *Quantidades Vendidas* e *Vendas em Dólar*, por exemplo.

A seguir, para melhor compreensão, serão descritas as etapas passo a passo.

8.2.1. Identificação do Processo a Modelar

O processo de negócios escolhido para ser modelado relaciona-se a uma rede de lojas de departamento. É um modelo para vendas no varejo que atua como intermediário entre os fabricantes de produtos e seus consumidores. Sua modelagem dimensional permitirá que sejam compreensíveis as tabelas de fatos e de dimensões.

Neste negócio, cada uma das lojas da rede é composta, como é típico, por uma grande variedade de departamentos, que inclui aqueles de calçados, roupas, perfumes, cama, mesa e banho, eletrodomésticos, utensílios, brinquedos, moda infantil e decoração. As lojas estão espalhadas em três estados e comercializam cerca de 60 mil produtos individuais.

Os itens são chamados de unidades de estoque ou SKUs (Stock Keeping Units) e a grande maioria provêm de fornecedores externos. Cada item apresenta um código de

barra impresso em sua embalagem, que é denominado Código Universal de Produto ou UPCs (Universal Product Code). Estes códigos representam o mesmo grão que SKUs individuais, de onde se conclui que cada variação de embalagem de um produto possui um UPC diferente, portanto, uma SKU também diferente. Os produtos restantes, ou seja, aqueles fabricados por terceiros, mas com a marca da loja, não possuem códigos UPCs reconhecidos em âmbito nacional. Entretanto, a própria loja ou unidade da rede deve atribuir um número SKU a esses produtos. Um dos maiores problemas com que se deparam os administradores de bancos de dados é achar uma chave mestra capaz de manipular tanto os UPCs quanto os SKUs. Assim, volta-se ao tema de atribuição de uma chave mestra às dimensões de nosso modelo.

Como qualquer negócio moderno e altamente automatizado, colam-se etiquetas para muitos itens de diversos setores. E embora os códigos de barras não sejam UPCs autênticas, certamente são números SKU. Não restando portanto, produtos a serem identificados em listas de controle.

Decisões administrativas mais significativas para qualquer organização relacionam-se aos lucros, promoções e política de preços. Por isso, a análise de suas operações poderá ser visualizada através dos dados do Data Warehouse. O lucro resulta do estabelecimento de uma margem máxima em termos financeiros, enquanto se reduz os custos de aquisição e os custos indiretos dos produtos, procurando o oferecimento desses últimos ao maior número possível de clientes, por meio de promoções e de uma política de preços altamente competitiva. Os dados que ajudam na análise do negócio em questão podem ser coletados nos diversos pontos de vendas de cada departamento no interior da loja, onde os clientes efetuam suas compras, assim como nos fundos da mesma, que é precisamente onde os fornecedores fazem suas entregas e se obtém o controle dos estoques. Estes dados podem contribuir para análises em torno da maximização do lucro, assim como com a rotatividade dos estoques e política de preços e promoções que tanto afetam o desempenho econômico/financeiro da empresa.

A seguir, tendo sido feita a descrição do negócio a modelar, passa-se ao passo seguinte, que se refere à escolha dos grãos da tabela que servirão para a tabela de fato e tabelas de dimensão.

8.2.2. Determinação do Grão das Tabelas

Tanto os Data Warehouses crescem em sofisticação e complexidade, que vão necessitando de organização em níveis cada vez mais altos. Durante muito tempo utilizou-se a classe como fator de organização para aplicativos como demostrativo da granularidade dos dados em nível de organização. Mas a definição e as características de classe mostraram-se de pouco alcance sobre aplicações cada vez maiores e embora muitos metodologistas tenham tentado identificar outros tipos de níveis básicos para tabelas dimensionais.

A granularidade é o fator que representa cada processo, em seu nível mínimo, na tabela de fatos. É também o que determina as dimensões do banco de dados usados pelo Data Warehouse. Através da granularidade é possível identificar a categoria ou o assunto de que tratam os dados e agrupá-los segundo suas afinidades, visando melhorar as respostas das consultas. A granularidade ocorre em todas as tabelas do modelo e à medida que cresce em detalhes, crescem visivelmente os efeitos da análise sobre os dados.

Como exemplo de granularidade, vai ser designada a **Dimensão Tempo.** As questões mais freqüentes feitas aos usuários sobre o nível de detalhe ou grão desta tabela são:

- *A consulta dos dados é feita ao dia, semana, mês ou quadrimestre?*
- *Deve-se distinguir dois dias diferentes da semana na análise dos dados? Por exemplo, quintas-feiras de sábados?*
- *É necessário consultar baixas no estoque de itens em promoção, durante o período em que a mesma é realizada?*
- *Qual o instantâneo de período suficiente para realizar o objetivo da consulta?(Dia, semana, mês, quadrimestre)*

Para designar a **Dimensão Produto**, tem-se:

- *Qual o nível de rastreamento do produto?*

1. *SKU*
2. *Lote*

Para designar a **Dimensão Loja**

- *É necessário rastrear vendas em nível:*

1. *Região*
2. *Cidade*
3. *Bairro*

Para designar a **Dimensão Cliente**

- *É desejável rastrear produtos vendidos em nível cliente?*

A combinação dos níveis de detalhe de cada uma das dimensões determina a granularidade da tabela de fatos. Então, dependendo do banco de dados, algumas vezes é necessário manter uma tabela de transações individuais ou um instantâneo de caráter cumulativo ou mensal, ou até mesmo combinações desses dois itens em tabelas de fato separadas.

Ressalta-se que uma definição cuidadosa do grão pode determinar as dimensões primárias na tabela de fatos. Em geral, é possível adicionar outras dimensões ao grão básico da tabela de fatos, sendo que estas dimensões adicionais devem produzir um

único valor para cada combinação das dimensões primárias. Entretanto, se for constatado que uma dimensão acrescentada é capaz de violar o grão, gerando registros adicionais, então a definição do grão deve ser revisada para acomodar esta dimensão adicional.

No presente caso, as tabelas de dimensão em utilização pela cadeia de lojas de departamentos que darão o nível de granularidade ao Data Warehouse são descritas logo a seguir :

- **Produto_id**
- **Tempo_id**
- **Cliente_id**
- **Promoção_id**
- **Loja_id**
- **Loja_Vendas**
- **Loja_Custo**
- **Unidade_Vendas**

Abaixo, serão feitas descrições minuciosas da composição das tabelas de fato e de dimensão, suas características e seus conteúdos.

8.2.3. Composição das Tabelas

Criar um Data Warehouse requer que se veja seu propósito de organizar vastos volumes de dados estáveis a fim de tornar mais fácil sua recuperação e análise. Da organização do DW depende seu acesso rápido às informações para análise e relatório de negócios. Por essa razão, o modelo dimensional mostra-se mais adequado às consultas de dados sumarizados e/ou disponíveis em grandes volumes, além, é claro, de tornar mais simples os esquemas projetados pelos analistas para tal fim.

Determinado qual o modelo a ser usado no banco de dados que dará suporte ao Data Warehouse da loja de departamentos e recaída a escolha sobre a modelagem dimensional dos dados passa-se à descrição das tabelas que lhe servirão de base.

Contrariamente ao modelo relacional que embora seja, freqüentemente, utilizado para se criar um único e complexo modelo para todos os processos da organização, o modelo dimensional cria a individualização das tabelas para detalhar os processos de negócios. E enquanto os primeiros mostram-se eficientes em transações do tipo on-line por possuir alto nível de normalização de dados, que é característica de utilizações maciças de sistemas OLTP, no modelo dimensional os dados concernentes às vendas podem ter modelagem diferente daqueles ligados ao inventário ou aos clientes, porque sua principal preocupação é a análise dos dados.

No modelo de esquema Star (Estrela), cada tabela de dimensão possui uma chave primária única a ligá-la à tabela de fato. Em esquemas Snowflake, por outro lado, as dimensões são decompostas em múltiplas tabelas com outras tabelas de dimensão subordinada que se juntarão à tabela de dimensão primária em vez de se juntar à tabela de fato, como ocorre no modelo anterior, o que a torna de um certo modo, inviável ao nosso projeto. É por esta razão que na maioria dos projetos, o esquema Star é preferível ao Snowflake, porque envolve um número menor de tabelas de dimensão e a busca se torna menos complexa. As tabelas de fato e de dimensão serão descritas em seguida:

- **Tabelas de Fato**

Cada Data Warehouse ou data mart inclui uma ou mais tabelas de fato centralizada e pertence a um esquema Star, que vai capturar os dados que medem as operações da organização. Uma tabela de fato deve conter eventos tais como as vendas, que registram cada uma das informações ou expedições, até mesmo de organizações sem fins lucrativos. Normalmente, as tabelas de fato contém um grande número de linhas, algumas em números que chegam à casa das centenas de milhares de registros, por guardarem informações de vários anos de história, no caso de grandes organizações. Uma característica chave de uma tabela de fato é que ela contém fatos numéricos que podem ser sumarizados a fim de prover informação sobre as operações da organização. Cada tabela também inclui um índice que contém, como chaves estrangeiras, as chaves primárias de tabelas de dimensões relacionadas, que por sua vez contém os atributos dos registros dos fatos. As tabelas de fato não devem conter informações descritivas ou

dados que não sejam numéricos, além dos campos com índices que relatam fatos com suas respectivas tabelas de dimensão.

Um exemplo da tabela de fato **Vendas_Fato_2001** da rede de lojas de departamentos que contém as seguintes colunas:

Fato	Descrição
produto_id	Chave Estrangeira para a tabela de dimensão **produto**.
tempo_id	Chave Estrangeira para a tabela de dimensão **tempo_por_dia**.
cliente_id	Chave Estrangeira para a tabela de dimensão **cliente**.
promoção_id	Chave Estrangeira para a tabela de dimensão **promoção**.
loja_id	Chave Estrangeira para a tabela de dimensão **loja**.
loja_vendas	Coluna geral contendo o valor da venda.
loja_custo	Coluna geral contendo o custo (para a loja) da venda.
unidade_vendas	Coluna numérica contendo a quantidade vendida.

Figura 8.2.3.1. Tabela de fato contendo as chaves para tabelas de dimensão

Nesta coluna de fato, cada entrada representa a venda de um produto específico, em um dia específico, para um cliente específico, de acordo com uma promoção específica em uma loja específica. As medidas de negócio capturadas são os valores correspondentes à venda, aos custos da venda para a loja, além das quantidades vendidas.

É preciso ressaltar que as medidas incluídas na tabela de fato são números aditivos. Esses números aditivos permitem uma informação sumária a ser obtida através da soma de várias quantidades da medida, tais como as vendas de um determinado item, em um grupo de lojas, para um período particular de tempo. Medidas não aditivas tais como os valores do inventário podem ser usadas na coluna de fato, desde que sejam usadas também, técnicas de sumarização que lhes sejam adequadas.

- **Tabelas de Dimensão**

Tabelas de Dimensão contém atributos que descrevem fatos registrados na tabela de fato. Alguns desses atributos provêem informações descritivas, enquanto outros são usados para especificar como os dados da tabela de fato devem ser sumarizados a fim de

oferecer informações úteis para os analistas. As tabelas de dimensão possuem hierarquias de atributos que ajudam na sumarização. Por exemplo, a dimensão que contém informações sobre produto deve possuir uma hierarquia que separa produtos por categorias, tais como roupas, produtos de beleza e brinquedos, entre outro itens, com uma sub-divisão de cada uma dessas categorias até que seja alcançado o nível de unidade de estoque ou SKU (Stock Keeping Unit).

A modelagem dimensional produz tabelas de dimensão nas quais são contidos atributos de fatos que são independentes de outras dimensões. A tabela de dimensão **Cliente** contém informações relacionadas somente aos clientes, assim como a dimensão **Loja** e a dimensão **Produto** vão conter somente informações relacionadas ao tema.

As consultas utilizam atributos em dimensões a fim de especificar a visualização das informações do fato. Se fosse necessário saber qual o custo das blusas femininas de seda vendidas na região norte do estado de Santa Catarina, teríamos uma consulta que envolve ao mesmo tempo as dimensões **Produto**, **Loja** e **Tempo**. Consultas subseqüentes devem "escorrer" até uma ou mais dimensões visando examinar mais detalhadamente os dados à medida que as tabelas são usadas para se especificar como uma medida (custo) contida na tabela fato está para a sumarização.

As colunas de uma tabela de dimensão podem ser usadas para categorizar informação em níveis hierárquicos. Abaixo a amostra da tabela para a rede de lojas de departamentos que inclui algumas colunas previamente escolhidas para especificar os níveis de hierarquia:

Coluna	Descrição
Loja_país	Especifica o país no qual a loja está localizada. Este é o nível de hieraquria do país.
Loja_estado	Especifica o estado no qual a loja está localizada. Este é o nível de hierarquia estado.
Loja_cidade	Especifica a cidade no qual a loja está localizada. Este é o nível de hieraquria cidade.
Loja_id	Especifica individualmente a loja. Este é o nível mais baixo de hierarquia. Este campo contém a chave primária da dimensão loja e é usado para juntar a tabela de dimensão à tabela de fato.
Loja_nome	Especifica o nome da loja. Os valores dessa coluna são usados para identificar a loja, de forma legível, para os usuários.

Figura 8.2.3.2. Tabela de dimensão

Outras colunas não mostradas aqui, podem prover informações com atributos adicionais. Entretanto, considera-se que haja uma estimativa do tamanho básico do banco de dados projetado inicialmente, pelos analistas do projeto. É preciso ressaltar que as tabelas de dimensão e dos índices associados a elas são, normalmente, menores do que a tabela de fatos e seus índices.

Como visto anteriormente, cada modelo de tabela, por assim dizer, captura fatos para a sua **Tabela de Fatos** e os atributos desses fatos são capturados para as **Tabelas de Dimensão** que estão ligadas à primeira. Normalmente são obtidos modelos com esquemas do tipo Star que tem se provado efetivo e eficaz nos projetos de Data Warehouse.

O modelo dimensional organiza as informações em estruturas que correspondem, freqüentemente, ao modo como os analistas preferem que seja a consulta de dados no DW. Por exemplo, a questão: "*Como estão as vendas de blusas de lã nas lojas da região oeste do estado de Santa Catarina no terceiro quadrimestre do ano de 2001?*", vai representar o uso de três dimensões a saber (produto, região e tempo) a fim de especificar as informações a serem sumarizadas. É um modelo dimensional simples de informações sobre vendas, mas deve incluir uma tabela de fato chamada **Fato Vendas**,

que irá conter um registro para cada linha de item vendido, a localização e data da venda efetuada, capturando assim, as informações das tabelas de dimensão chamadas **Produto**, **Região** e **Tempo**.

O registro da variedade de informações sobre vendas deve incluir outras informações importantes tais como aquelas sobre clientes (caso haja um cartão com seu nome e/ou número de registro) e a loja onde ocorreu a transação, além do período de tempo, a data e detalhes sobre o produto vendido. Cada uma dessas categorias de informação deverá ser organizada em sua própria dimensão. Ou seja, a informação sobre o cliente será colocada na tabela de dimensão **Clientes**, enquanto as informações sobre a loja e o período de tempo em que ocorreu a venda do item, poderão ser localizadas nas tabelas de dimensão **Loja** e **Tempo**, respectivamente.

8.3. Fatos Mensuráveis

Fatos mensuráveis estão ligados ao processo de popular o DataWarehouse e seus data marts a partir de fontes externas e operacionais. Construir um modelo inicial dimensional serve para identificar os elementos candidatos a popular o DW e facilitar a análise pelos usuários. Como os requerimentos de análise de dados em geral são não-lineares e sabendo-se que as relações entre eles produzem diferentes visões da realidade, é necessário estabelecer uma abordagem que capture e apresente estes elementos em sua melhor disposição.

A abordagem para a efetivação dos elementos que povoam as tabelas de fato e dimensão procura determinar em primeiro lugar os itens mensuráveis, em seguida as dimensões associadas a eles e por último, os fatos. Esta é uma abordagem orientada à busca, pois tende a fluir naturalmente quando os requerimentos de análise dos usuários vão ao encontro dos dados colocados à sua disposição.

Elementos que se mostram bons candidatos a se tornarem fatos mensuráveis são numéricos, envolvidos em cálculos de agregação e profundamente voltados para a

análise, que é o requisito principal para popular o Data Warehouse. Mesmo assim, nem todo atributo numérico pode vir a ser válido como fato mensurável, pois para tanto, é preciso que possua propriedades peculiares como a capacidade de associação com várias dimensões e um nível de granularidade capaz de determinar a melhor combinação no armazenamento de seus detalhes, em todas as dimensões envolvidas. Isto implica em um sério aumento do Data Warehouse, o que além de torná-lo mais complexo, também pode causar um grande impacto em termos de performance. Portanto, fatos mensuráveis podem influenciar muito o Data Warehouse, mas a recíproca também é verdadeira.

Um bom exemplo de fato mensurável é a quantidade de mercadorias vendidas em um determinado período de tempo, em uma das lojas da rede. Tem-se aí quatro tabelas de dimensão - **Produto**, **Loja**, **Tempo**, **Unidade_Venda** - onde tal elemento numérico pode ocorrer e uma tabela de fato - **Vendas**.

Analisando-se o contexto da busca, nota-se que o fato numérico a que se refere o exemplo acima (quantidade de mercadorias vendidas) é constante em todas as cinco tabelas, o que confirma ao mesmo tempo, os seus predicados de associação e granularidade.

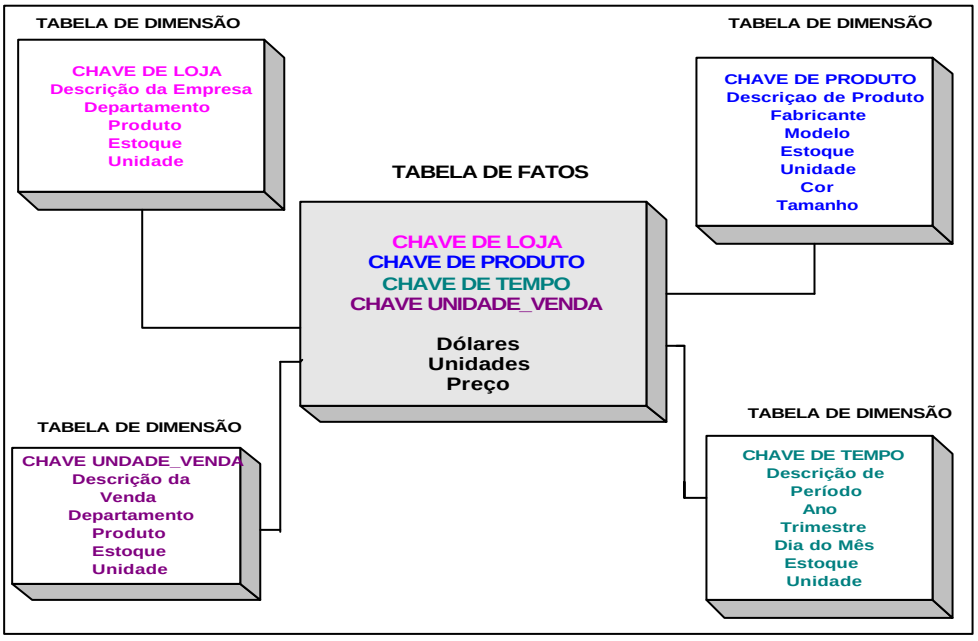

Figura 8.2.4. Tabela de fato e tabelas de dimensão

8.4. Geração de Chaves

Selecionar chaves em um Data Warehouse é uma escolha difícil, pois envolve uma espécie de negociação entre a performance e o gerenciamento que se aplica à muitas dimensões. Há dois tipos de chaves em no modelo dimensional, as chaves primárias (Primary Key) e as chaves externas (Foreign Key). Em geral, as chaves escolhidas para as dimensões correspondem às chaves externas do fato.

Chave primária constitui-se de uma ou mais colunas da tabela, que procuram identificar somente aquela linha dentro da tabela, assegurando também a unicidade da chave. A razão pela qual se especifica uma chave primária é garantir a integridade da tabela, além de poderem ser usadas em *joins* (junções), para relacionar uma tabela às outras. Por exemplo, todas as tabelas de dimensão, **Prod_id**, **Temp_id**, **Clie_id**, **Prom_id** e **Loj_id** são chaves primárias das tabelas Produto, Tempo, Cliente, Promoção e Loja. Já a tabela de fato **Loja_Vendas** tem sua chave primária composta pelas colunas **Prod_id**, **Temp_id**, **Clie_id**, **Prom_id** e **Loj_id**.

Tabela de Fato

Loja_Vendas

Coluna	Descrição
Prod_id	Este campo contém a chave primária da dimensão **produto** e é usado para juntar a tabela de dimensão à tabela de fato **Loja_Vendas**.
Temp_id	Este campo contém a chave primária da dimensão **tempo** e é usado para juntar a tabela de dimensão à tabela de fato **Loja_Vendas**.
Clie_id	Este campo contém a chave primária da dimensão **cliente** e é usado para juntar a tabela de dimensão à tabela de fato **Loja_Vendas**.
Prom_id	Este campo contém a chave primária da dimensão **promoção** e é usado para juntar a tabela de dimensão à tabela de fato **Loja_Vendas**.
Loj_id	Este campo contém a chave primária da dimensão **loja** e é usado para juntar a tabela de dimensão à tabela de fato **Loja_Vendas**.

Figura 8.4. Tabela de fato contendo dimensões, com suas chaves primárias.

Quando se cria uma chave primária em um banco de dados, por exemplo do tipo SQL Server, aparece uma chave na coluna que antecede aquela cujo nome estará ligado à chave primária. Outra observação importante é que chaves primárias não podem conter valores nulos (Null).

Chave externa é uma coluna ou associação de colunas usada para estabelecer ligação entre duas tabelas através de campo ou campos comuns e usada para manter a integridade referencial entre as duas tabelas. Se tivermos uma tabela com os nomes de funcionários da loja, onde existe um campo chamado Departamento. Este campo é associado ao campo **Dep_id** de uma tabela chamada Dept, a qual possui dados sobre todos os departamentos da loja. Caso não existisse uma chave externa para este campo, uma linha contendo o nome de um departamento, se existirem mais de um funcionário com o nome daquele departamento especificado em seus registros.

Escolher os campos para os quais serão geradas chaves primárias não é uma tarefa simples. Em princípio, qualquer campo pode servir perfeitamente para gerar a chave primária, mas é imperativo nesta escolha, que se leve em consideração o principal significado de identificação das dimensões, ou seja, que aquela linha que está sendo identificada por uma chave primária não será objeto de identificação do mesmo tipo por parte de outras tabelas, resguardando-se assim a característica de unicidade da tabela e da chave.

8.5. Criação de Índices

Índices são tipos especiais de arquivos empregados em associações com tabelas. Os índices têm um papel de extrema importância no desempenho do Data Warehouse, assim como em todos os bancos de dados relacionais e dimensionais, pois podem também ser úteis em consultas especializadas, acelerando o processo de acesso a um dado registro ou grupo de registros. Criar um índice permite que, como em um livro, não se perca tempo folheando as páginas em busca de um determinado assunto, ao contrário, pela consulta do índice, vai-se diretamente ao ponto de interesse. Vista a sua utilidade, cada tabela de dimensão, portanto, deve ser indexada através de sua chave primária.

A coluna de fatos deve ser indexada através da composição de chaves estrangeiras das tabelas de dimensão. Estes são os índices primários necessários à maioria das aplicações do Data Warehouse por causa da simplicidade requerida pelo esquema Star. Requerimentos de consultas e relatórios especiais podem indicar a necessidade de índices adicionais.

Considerando-se o caso da rede de lojas de departamentos, a tabela Cliente tem um índice baseado na coluna **Clien_id**:

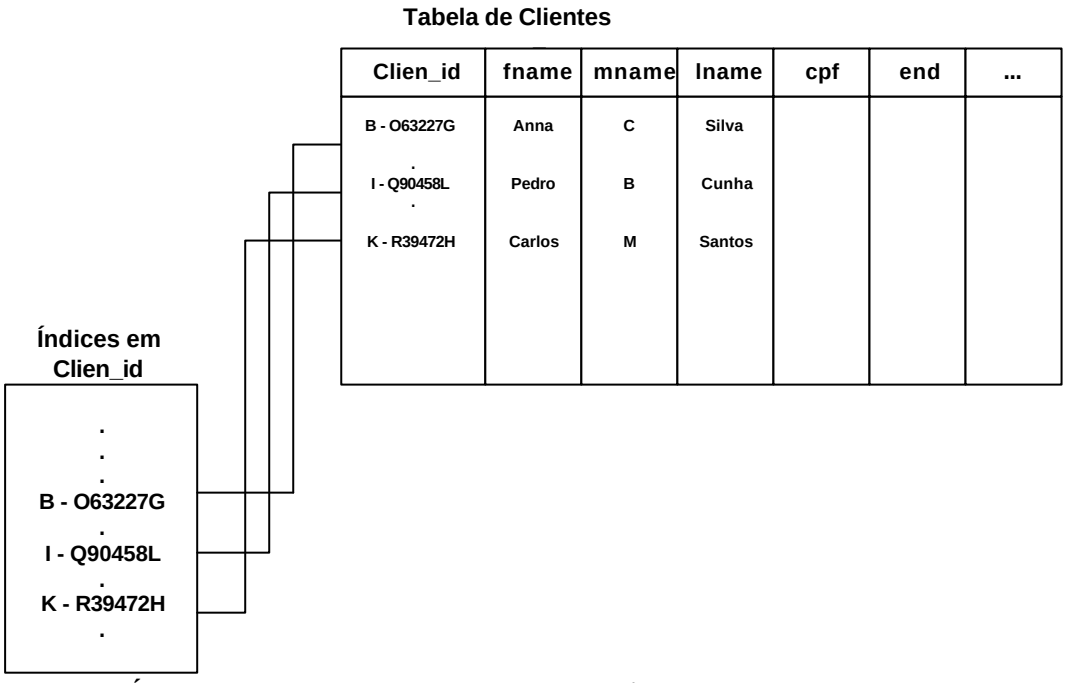

Figura 8.5. Índices criados a partir da tabela de dimensão Clientes

Ao se executar uma consulta na tabela Cliente, o servidor procura e detecta a coluna-chave e pesquisa no índice, que contém uma cópia do conteúdo da coluna **Clien_id** e seu endereço, o da linha, dentro da tabela. Há índices que são criados automaticamente, como é o caso das colunas que possuem chaves primárias e há os que necessitam ser criados manualmente. O ideal é que cada coluna da tabela possua um índice associado, melhorando o desempenho na busca, mas isto ocupa espaço em disco com uma função nem sempre utilizada.

A criação de chaves pode se basear sobre os propósitos de análise ou simplesmente sobre a criação de um nível desejado de granularidade. Questão que envolve tanto a performance quanto a manutenção, mas não deixa de ser um fato a ser documentado no metadado, seja por motivos técnicos, seja por motivos administrativos.

8.6. Visões de Tabelas

Visão é um conjunto de instruções que retornam ao usuário um conjunto de dados. Uma visão é, basicamente, uma tabela virtual com o conteúdo definido por uma consulta ao banco de dados. Porém, não é uma tabela física e se apoia em duas ou mais tabelas. Por isto, a visão permite que os usuários possam ter acesso a dados de diversas tabelas, tantas quantas compuserem a visão. A única restrição seria que as tabelas que formam a visão tenham afinidade entre seus dados.

Abaixo, a figura de uma visão composta por duas tabelas:

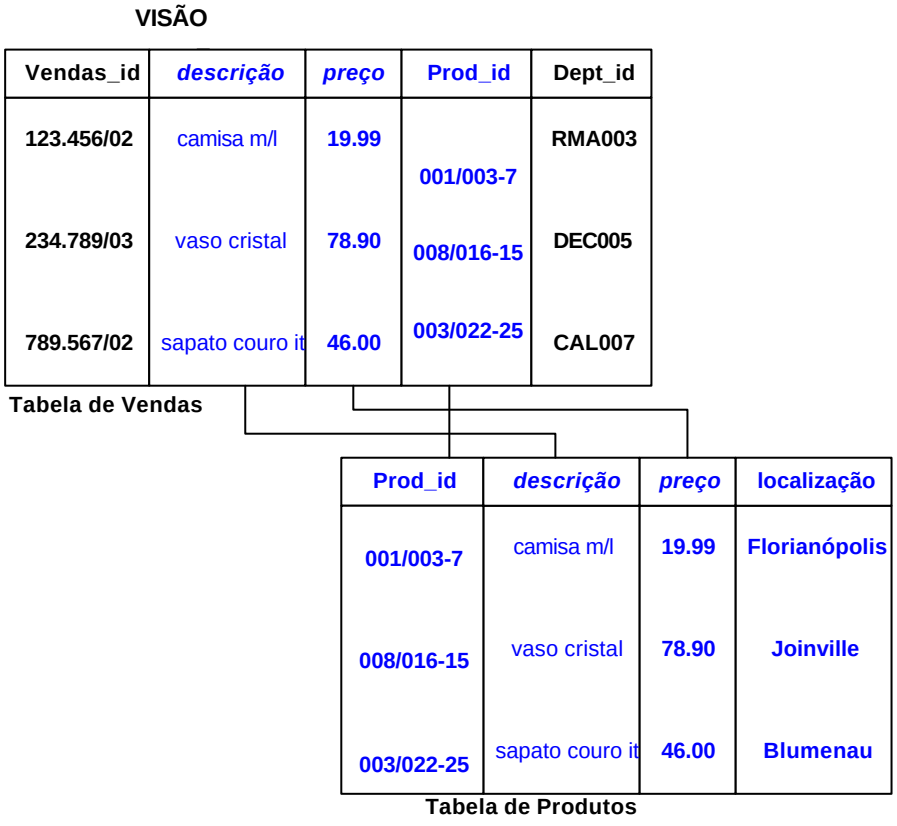

Figura 8.6. Visão da tabela de dimensão Produto

A visão das vendas é composta pelos dados das tabelas de Vendas e da tabela de Produtos, que oferecem informações diversas tais como a descrição da venda do

produto, código e preço. Com as visões pode-se navegar sobre várias tabelas de dimensão que integram um fato e que, juntas, conseguem compor um quadro de uma determinada situação, no tempo.

8.7. Desenvolvimento de Diagramas

Diagramas são a representação dos componentes dos bancos de dados que capacitam os usuários a visualizar as tabelas e seus relacionamentos. Quer dizer, as tabelas, índices e visões armazenadas pelo banco de dados dimensional do Data Warehouse podem ser vistos e manipulados com o uso de recursos *clicar e arrastar* e ainda permitem a interação com caixas de diálogo em execuções de diversas tarefas, como a alteração de características da base de dados, adição ou eliminação de tabelas, triggers, stored procedures, constraints ou relacionamentos, sem a necessidade de usar qualquer tipo de linguagem de programação. Mas é preciso dizer que cada tabela exibida pelo diagrama é tão somente uma referência à tabela que está armazenada, fisicamente, no banco de dados. Os diagramas podem ser criados em muitos números para uma mesma base de dados e, além disso, estão habilitados a realizar experimentos na estrutura do banco de dados sem executar realmente as modificações e se for o caso, executá-las, de fato. Neste caso, quando se efetuam modificações nas características de uma tabela, essas modificações passam a ser automaticamente refletidas nos demais diagramas em que esteja relacionada a tabela modificada. Mas as alterações feitas no diagramas só passam a ser incorporadas à tabela, se o mesmo for salvo.

Na próxima página está o exemplo de um diagrama da loja da rede de departamentos.

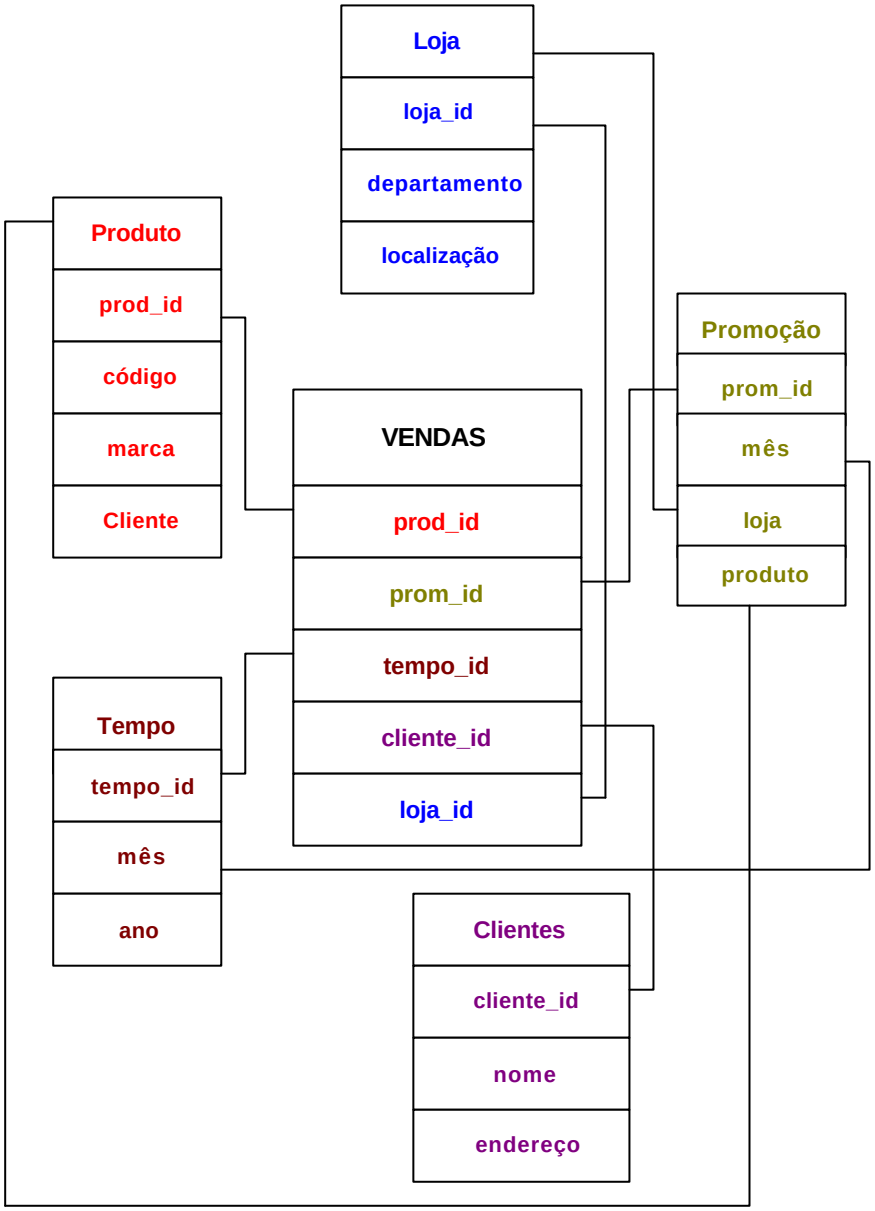

Figura 8.7. Diagrama mostrando o relacionamento entre tabelas

8.8. Resultados

O resultado do presente trabalho procurou ir de encontro à teoria referenciada mais adiante. Muitos dos conceitos são facilmente aplicáveis, sem dúvida dada a sua simplicidade e consistência.

Conceitos de dimensões, fato, tabelas, visões e diagramas, entre outras coisas, facilitaram a organização do trabalho. E embora este não tenha tido uma aplicação prática destes preceitos, é certo que muitas das ferramentas de bancos de dados disponíveis no mercado foram concebidas e construídas sobre eles.

É evidente que muitos passos foram simplificados, visando o melhor entendimento dos conceitos apresentados. É preciso dizer também, que a modelagem de um Data Warehouse não é tão carente de complexidade quanto possa parecer, pois envolve algum tempo de estudo da própria organização, sua missão, objetivos gerais e específicos, sua cultura, aspectos humanos e, principalmente, o seu papel no mercado. A partir deste estudo, parte-se, então para uma análise mais especifica do ambiente interno da organização, agora, levando-se em consideração a sua necessidade de informação em suas atividades rotineiras e em seus processos de tomada de decisão.

Aqui, procurou-se demonstrar, dentro dos conceitos apresentados, uma forma simplificada de modelar um Data Warehouse. É claro que em se tratando de uma empresa fictícia, não se pode comprovar a efetividade do projeto. O que também não invalida o trabalho, pois é facilmente aplicável a sua teoria no caso de se utilizar uma boa ferramenta de construção de bancos de dados, com orientação para Data Warehouse e se possível, obtendo-se dados retirados do ambiente real de uma organização, o que sem dúvida traria melhores resultados ao trabalho. Mas visto que por diversos motivos, as empresas não oferecem a possibilidade de pesquisas deste caráter, fica a proposta com a modelagem do DW para futura implementação, já que é mesmo este o seu principal objetivo.

8.9. Discussão

Os dados voltados para Data Warehouse, por outro lado, requerem não somente sua extração e armazenamento, mas também seu refinamento, por dizerem respeito à

tomada de decisão, porque Data Warehouses, por definição, criam repositórios desnormalizados de dados, que levam os usuários à análises mais rápidas e inteligentes dos negócios.

A maioria das organizações tem a preocupação excessiva em relação ao armazenamento dos dados como se o Data Warehouse fosse apenas um depósito para estes últimos. Isto leva, com freqüência, ao descompasso em relação às suas necessidades de informação.

São inúmeras as empresas que investem tempo, equipamento, pessoal e recursos financeiros em sistemas de informações não lhes trazem qualquer benefício de ordem prática. Uma boa parte dos usuários destes sistemas permanecem à deriva quando se trata do bom uso dos dados que supostamente lhes estão disponíveis, pois como os referidos sistemas são construídos sem qualquer preocupação com os "consumidores" dos dados, estes tornam-se inúteis não só nos processos decisórios, mas também em trabalhos rotineiros como a apresentação de relatórios. É de se entender, portanto que a teoria em torno do Data Warehouse deve ser estudada e compreendida, antes mesmo que se pense em investir em equipamentos que tenham este fim.

O Data Warehouse em si não é uma ferramenta tão complexa. Complexos são os seus requerimentos de dados. Compreende-los é uma árdua tarefa que cabe não só aos analistas, mas também aos usuários. Todo o resto, modelos, esquemas, tabelas, visões e tudo mais, são apenas elementos que podem ser aprendidos e usados, desde que se absorva o Data Warehouse como uma espécie de cultura da empresa e não apenas como um sistema de informações que funcionará por si mesmo, tão logo seja adquirido um bom software.

8.10. Conclusões

A razão pela qual, um sistema operacional que tem suas aplicações voltadas para o fluxo diário de trabalho da organização e armazena vastos volumes de informações detalhadas é inadequado ao processo de tomada de decisão, ainda é um paradoxo.

Muitos sistemas operacionais se apoiam em transações recorrentes de dados armazenados em tabelas relacionadas e seguindo esquemas normalizados que, embora

limitem a redundância e promovam a saída de informações, são capazes de obstruir o processo decisório, isto porque dados normalizados são reconhecidamente difíceis de analisar através de consultas do tipo ad hoc.

Considerando-se ainda que uma organização, normalmente, não possui pessoal com tempo, habilidade ou ferramentas para fazer buscas a partir de sistemas operacionais e o suporte para tarefas deste gênero e que são os analistas e programadores os intermediários que escrevem as buscas e os relatórios dos programas, conclui-se que sem um bom projeto de DW, os usuários são deixados para “caçar” dados, enquanto aos programadores resta manterem-se alertas com o fluxo constante de novas requisições.

Por outro lado, ao permitir que os usuários adquiram e analisem dados com facilidade, o Data Warehouse torna poderosos os responsáveis pela tomada de decisão, deixando programadores livres para enfocar objetivos mais estratégicos para as informações. Data Warehouses, portanto, necessitam de planejamento e projeto adequados às necessidades de informação da empresa. Quando devidamente construídos, eles podem ser fontes de grandes vantagens competitivas, as quais não existem quando se trata de simples armazenamento de dados.

Muitos especialistas se dedicam a desvendar os mistérios do Data Warehouse para a empresas. Inúmeras são as obras são publicadas e anunciadas em todos os tipos de mídia. Entretanto, o mistério em torno do correto uso das informações persiste. O mas importante a se saber sobre este fato é que o Data Warehouse, apesar de ser um conceito universal em tanto que teoria, deve ser, antes de mais nada, para a organização que deseja implementá-lo, espécie de procedimento diário, daqueles que todos os funcionários podem e devem assimilar como parte de sua rotina de trabalho.

CONCLUSÃO

Mercados financeiros têm sido não apenas palcos de grandes disputas entre organizações e suas adversárias, como também têm sido a mola que impulsiona novas e importantes descobertas tecnológicas.

A maior parte das ferramentas orientadas a negócios são voltadas para melhorar sistemas apoiados em informações. Assim foram criadas as bases para o Data Warehouse. Assim também foram incrementados e tornados muito mais possantes bancos de dados de todos os tipos que possuem tal finalidade.

A proposta deste trabalho foi aplicar uma metodologia na construção de um Data Warehouse que representasse e sintetizasse as características mais relevantes da ferramenta. Para isso, foram consultadas obras de diversos pesquisadores da área e baseado em suas experiências (práticas, em alguns casos), formou-se uma metodologia capaz de dar feições a um modelo de Data Warehouse para uma empresa varejista, com necessidades bastante comuns até mesmo a diferentes ramos.

Através do presente trabalho, pode-se verificar a aplicabilidade de alguns conceitos formulados pelos melhores autores neste segmento de pesquisa e embora tenha sido necessário conhecimento prévio sobre Bancos de Dados e seus modelos, foi relativamente fácil a aplicação de suas teorias, assim como aquelas relacionadas aos Data Warehouse.

Tendo-se consciência da limitação do período de tempo para a construção de um Data Warehouse mais abrangente, foi idealizado um modelo que contivesse as áreas mais populares em termos de informações, também porque estas são, em princípio, as que se relacionam de modo mais integrado com as outras áreas da empresa.

Em termos de projeto, espera-se que possa contribuir para futuras pesquisas na área que se anuncia uma das mais promissoras tanto em se tratando da tecnologia desenvolvida quanto se apresenta uma daquelas com as maiores demandas em torno da dos sistemas de informações. Portanto, fica o registro do valor significativo que ferramentas de integração e análise de dados tem para incrementar a competitividade das empresas no mundo atual.

REFERÊNCIAS BIBLIOGRÁFICAS

[ANNES] R. **Parallel Architecture For Natural Language Processing**. New York: VecPar. 2000.

[BERTALANFFY] L. V. **General System Theory: Foundation, Development, Applications.** São Paulo: Atlas. 1976.

[BIO] S. R. **Sistemas de Informação – Um Enfoque Gerencial**. São Paulo: Atlas. 1996.

[BOAR] B. **Understanding Data Warehousing Strategically**. New York: NCR's Communication Industry Line of Business. 1997.

[CHAUDHURI] S.; UMESHWAR, D. **An Overview of Data Warehousing and OLAP Technology**. London: ACM Sigmod Record. 1997.

[COOD] E.F. **Providing OLAP to User-Analysts: An IT Mandate**. 1993. Disponível em: <http://www.hyperionsolution.com>. Acesso em: 20 jun. 2001.

[DILLY] R. **Data Mining – An Introduction**. 1996. Disponível em: <http://www.dinfocenter.edu/~dilly>. Acesso em: 11 mar. 2001.

[FINKELSTEIN] C. **Business Re-Engineering And The Internet: Transforming Business for a Connected World**. Sidney: Information Engineering Services Pty Ltd. 1998.

[FINKELSTEIN] C. **Information Engineering** : Essentital Strategies. Sidney: Information Engineering Services Pty Ltd. 1999.

[FINKELSTEIN] C. **Information Engineering Services**. Sidney: Information Engineering Services Pty Ltd. 2000.

[FIRESTONE] J.M. **Evaluating OLAP Alternatives**. Sidney: Information Engineering Services Pty Ltd. 1997.

[FRAWLEY] W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge Discovery in Databases: An Overview. **AI Magazine**, n.13, v.3. Apr. 1992.

[GUPTA] V. R. **An Introduction to Data Warehousing**. 1996. Disponível em: <http://www. dwinfocenter.edu/~dilly>. Acesso em: 13 abr. 2000.

[HAISTEN] M. Planning For a Data Warehouse. **Info DB**, n. 2, v. 9. Mar.1996.

[HAISTEN] M. Designing a Data Warehouse. **Info DB**, n. 4, v. 9. Mar.1996.

[HOLSHEMIE] L.; SIEBES, M. **Modeling na Information System**. New Jersey: Prentice Hall PTR.1994.

[HYPERION] WHITE PAPERS. **The Role of OLAP Server in Data Warehousing Solution**. 2000. Disponível em: <http://www.hyperionsolutions.com>. Acesso em: 21 mai. 2001.

[INMON] W. H. **What is a Data Warehouse**. New York: Prism Solution Inc. 1995.

[KELLY] F. **Implementing an Executive Information System**. New York: WEB Media Co. 1999.

[KIMBALL] R. **A Dimensional Modeling Manifesto**: DBMS and Internet Systems. 1996. Disponível em: <http://www.dwinfocenter.edu/~kimball>. Acesso em: 25 jul. 2001.

[KIMBALL] R. **Data Warehouse Toolkit**. São Paulo: Makron Books do Brasil.1998.

[KIMBALL] R. **Is ER Modeling Hazardous to DSS?.** 1997. Disponível em: <http://www.dwinfocenter.edu/~kimball>. Acesso em: 13 ago. 2001.

[KORZYBSKI] A. What is Metadata?. **Data Warehouse Tools Bulletin. Issue**, n. 3, v.5. Mar. 1996.

[KORTH] H. New Focal Points for Research in Database Systems. **ACM Computing Surveys**, n. 8, v. 4. Apr.1996.

[KORTH] H.; SILBERSCHATZ, A. **Sistema de Bancos de Dados**. São Paulo: Makron Books do Brasil. 1997.

[LAUDON] K. C.; LAUDON, J. P. **Management Information Systems – New Approaches to Organization & Technology.** New Jersey: Prentice Hall PRT. 1998.

[LITWIN] P. & REDDICK, G. Fundamentals of Relational Database Design. **Microsoft Access 2 Developer's Handbook. Sybex**. n.1, v.1.1994.

[MARTIN] J. **Engenharia da Informação**. Rio de Janeiro: Editora Campus.1991.

[MELENDEZ] R. **Prototipação de Sistemas de informação – Fundamentos, Técnicas e Metodologia**. Rio de Janeiro: LTC – Livros Técnicos e Científicos. 1990.

[ORR] K. **Data Warehouse Technology**. 1997. Disponível em: <http://www.kenorinstitute.edu/~orr. Acesso em: 20 mar. 2000.

[PENDSE] N. **What is OLAP?.** 1999. Disponível em : <http://www.dwinfocenter.edu/~pendse>. Acesso em: 15 abr. 2001.

[PERKINS] A. **Developing a Data Warehouse – The Enterprise Engineering Approach**.1996. Disponível em: <http://www.visiblesystems.com/whitepapers>. Acesso em: 10 fev. 2001.

[REZENDE] D.A. **Engenharia de Software Empresarial**. Rio de Janeiro: Brasport.1997.

[TANLER] R. **Intranet Data Warehouse**. Rio de Janeiro: IBPI Press. 1998.

[WHITE] C. A Technical Architecture for Datawarehousing. **InfoDB**. n. 3, v.5. Aug. 1995.